PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

DESIGN AND EVALUATION OF AN INTELLIGENT USER INTERFACE IN EVIDENCE BASED HEALTH CARE

IVANIA DONOSO GUZMÁN

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Advisor:
DENIS PARRA SANTANDER

Santiago de Chile, August 2017

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

DESIGN AND EVALUATION OF AN INTELLIGENT USER INTERFACE IN EVIDENCE BASED HEALTH CARE

IVANIA DONOSO GUZMÁN

Members of the Committee:
DENIS PARRA SANTANDER
HANS LÖBEL
GABRIEL RADA GIACAMAN
RODRIGO PASCUAL JIMÉNEZ

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Santiago de Chile, August 2017

Gratefully to my parents, sisters,
boyfriend and dog.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Denis Parra of the Department of Computer Science at Pontificia Universidad Catlica de Chile. He was always available to answer my questions or to guide me whenever I had a problem. He consistently steered me in the right the direction whenever he thought I needed it. He helped me to go two conferences during my master and

I would also like to thank who were involved in my thesis for this research project: Gabriel Rada and Daniel Prez. Without theirparticipation and input, the thesis could not have been successfully conducted. I would like to thank specially to Gonzalo Bravo, who helped me during the user study we conducted.

I would also like to acknowledge Phd. Hans Lobe of the Department of Computer Science at Pontificia Universidad Catlica de Chile as the second reader of this thesis, and I am gratefully indebted to his for his very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my parents, to my two sisters, my boyfriend and my dog for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Ivania Donoso Guzmán

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Evidence-based health care (EBHC) is an important practice of medicine which attempts to provide systematic scientific evidence to answer clinical questions. In this context, *Epistemonikos* (`www.epistemonikos.org`) is one of the first and most important online systems in the field, providing an interface that supports users on searching and filtering scientific articles for practicing EBHC. The system nowadays requires a large amount of expert human effort, where close to 500 physicians manually curate articles to be utilized in the platform. In order to scale up the large and continuous amount of data to keep the system updated, we introduce *EpistAid*, an interactive intelligent interface which supports clinicians in the process of curating documents for *Epistemonikos* within lists of papers called *evidence matrices*. We introduce the characteristics, design and algorithms of our solution, as well as a prototype implementation and a user study to show how our solution addresses the information overload problem in this area.

Keywords: information retrieval, evidence based health care, controllable user interfaces.

RESUMEN

La Medicina Basada en Evidencia (EBHC por sus siglas en inglés) es una importante práctica de la medicina que intenta proporcionar evidencia científica de forma sistemática para responder a preguntas clínicas. En este contexto, *Epistemonikos* (`www.epistemonikos.org`) es uno de los primeros y más importantes sistemas en línea en el campo, proporcionando una interfaz que apoya a los usuarios en la búsqueda y filtrado de artículos científicos para practicar EBHC. El sistema hoy en día requiere una gran cantidad de esfuerzo humano, donde cerca de 500 médicos manualmente revisan los artículos para ser utilizados en la plataforma. Con el fin de ampliar la cantidad de datos y para mantener el sistema actualizado, introducimos *EpistAid*, una interfaz inteligente interactiva que apoya a los médicos en el proceso de encontrar documentos para responder una pregunta médica. Presentamos las características, diseo y algoritmos de nuestra solución, así como una implementación de prototipo y un estudio de usuario para demostrar cmo nuestra solución aborda el problema de sobrecarga de informacin en esta área.

Palabras Claves: recuperación de información, medicina basada en evidencia, interfaces de usuario controlables.

1. INTRODUCTION

1.1. Evidence Based Health-Care

Evidence-Based Health Care (EBHC) is a medical practice approach that emphasizes the use of research evidence to justify a medical treatment. Sackett et al. defined it as *"the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients"* (Sackett et al., 1996). This medical approach classifies information according to their ability to create high-trust recommendations. EBHC has produced a large impact in the practice and teaching of medicine, since applying the knowledge gained from large clinical trials to patient care promotes consistency of treatment and optimal outcomes, in contrast to solely relying on habits or anecdotal cases (Sackett & Rosenberg, 1995).

Clinicians have established a process to design evidence based guidelines (Khan et al., 2003) as seen in figure 1.1. First, clinicians pose a question, then seek studies related to it, select documents that are relevant to the question, evaluate the selected documents according to their quality, then perform a meta-analysis to finally obtain conclusions. These conclusions are presented in an article called *Systematic Review* (SR). The articles used in the meta-analysis are called *Primary Studies* (PS). A PS can be any study design, qualitative or quantitative, where data is collected from individuals or groups of people.

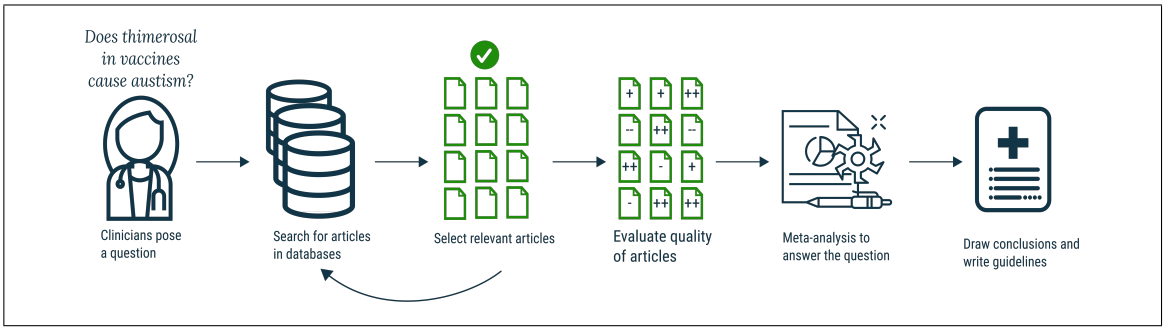

Figure 1.1. Process followed to answer a medical question.

The process of finding and selecting relevant documents is divided in several steps (figure 1.2). First researchers have to create a *search strategy* according to the question they established. A *search strategy* is a boolean query of keywords, for example `(vaccines OR vaccinate) AND thimerosal AND (children OR infants)`. It takes time to create them because physicians have to think all the possible ways to pose the question. This makes them very complex and prone to errors. They use the *search strategy* in several databases and gather all the articles they find. After that, clinicians perform *citation screening*, the process of manually reviewing titles and abstracts of articles to identify potentially eligible articles for inclusion in the review (Higgins JPT, 2011). After this process, only 2% to 15% of the articles remain (Wallace, Small, Brodley, Lau, & Trikalinos, 2010). Once they have selected the articles, they look for the full texts of these studies. This step is easy but bureaucratic. It includes, but is not limited to, looking for specific papers in databases, finding the correct version and paying for access. For this reason, it can take some time to get all articles. Clinicians review the full texts to finally keep only the articles that have high methodology quality. The meta-analysis is done using the information available in those articles only.

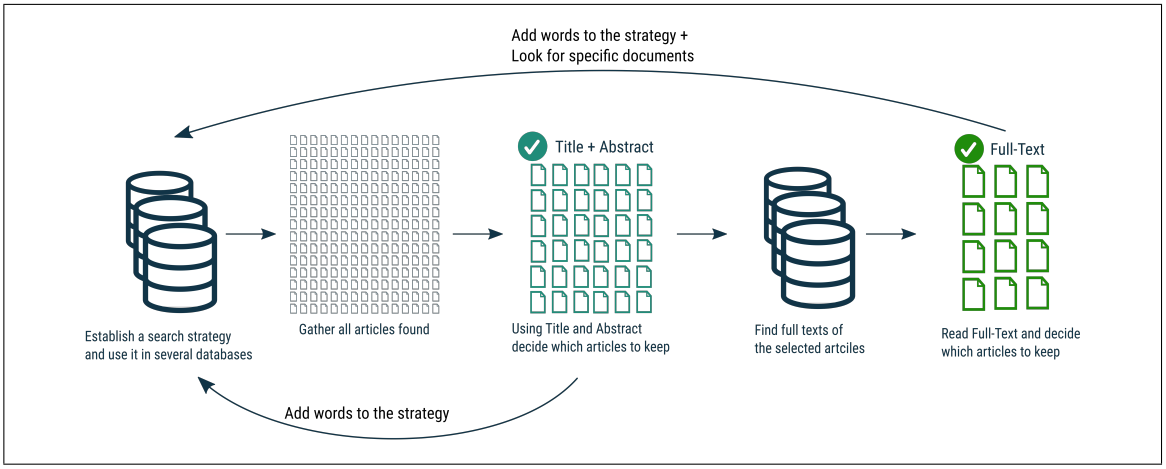

Figure 1.2. Process of finding relevant evidence

In EBCH, having all the relevant documents for a research question is critical (O Mara-Eves et al., 2015). It can happen that if they miss relevant evidence, the answer they

give to the question can be wrong or biased. For this reason the process of figure 1.2 is repeated iteratively and many people participate on it which make it very time consuming (Cohen et al., 2010; Bekhuis et al., 2014; Miwa et al., 2014; Elliott et al., 2014). This situation can be problematic because in practice, health-care related decisions must be made quickly (Thomas et al., 2011). Moreover, with the explosion of scientific knowledge being published, it is difficult for clinicians to stay updated on the latest best medical practices (Cohen et al., 2010; Elliott et al., 2014).

In this context, some systems attempt to support clinicians in the process of collecting, organizing, and searching for scientific evidence such as Embase (Elsevier, 2016), Covidence (Adams et al., 2013), and *Epistemonikos* (Rada et al., 2013). In particular, *Epistemonikos* is a collaborative database which stores research articles that provide the best evidence according to the EBHC principles (Rada et al., 2013). Since the evidence comes from scientific literature, this information is collected from specialized online sites such as PubMed[1] and Cochrane[2] , among more than 20 other sources of scientific information[3]. In addition to collecting, indexing and classifying medical research articles that have the quality to be used in EBHC, *Epistemonikos* developed a new way to find evidence to answer a medical question.

1.2. Epistemonikos, a collaborative research database

To ensure that no relevant evidence will be missed, physicians check the references of the papers they classified as relevant. In the case of systematic reviews, they do citation screening on the articles that were used in the meta-analysis of that SR. The idea behind this action is that the sets of primary studies of used in the meta-analysis of two SR that answer similar question will have several articles in common (figure 1.3).

[1] `https://www.ncbi.nlm.nih.gov/pubmed/`
[2] `http://www.cochranelibrary.com/`
[3] `http://www.epistemonikos.org/en/about_us/methods`

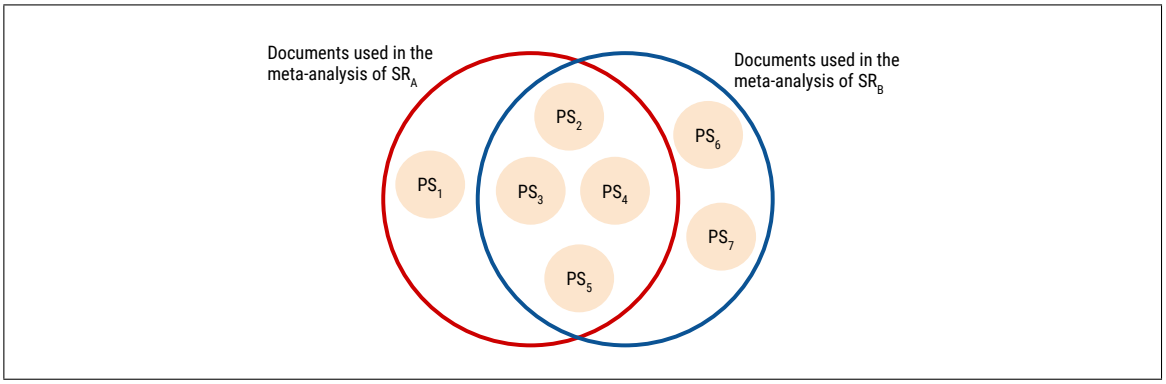

Figure 1.3. Finding more documents for systematic review A from systematic review B.

Epistemonikos used this fact to create a list of papers that could be relevant for a medical question. They call this list of papers an *Evidence Matrix*. The method for building the initial matrix M_0 is shown in figure 1.4. It starts with a user selecting a seed SR, that answers a similar clinical question. Using Breadth-First Search over the citation graph (Cormen, 2009), all the PS cited in the seed SR are added to the matrix (Level 1). Next, other SR in the database citing the papers in Level 1 are added (Level 2). Finally, other PS cited in the Level 2 are added as additional primary studies (Level 3).

Users remove papers from this *Evidence Matrix* doing citation screening: they check the title, abstract and other metadata to decide whether to keep or to remove an article. They do this process using the interface of figure 1.5.

After users have removed all non-relevant articles from the matrix, they can inspect a visual representation of an Evidence Matrix (figure 1.6). In the matrix, rows represent systematic reviews and columns are primary studies which have been cited in those systematic reviews.

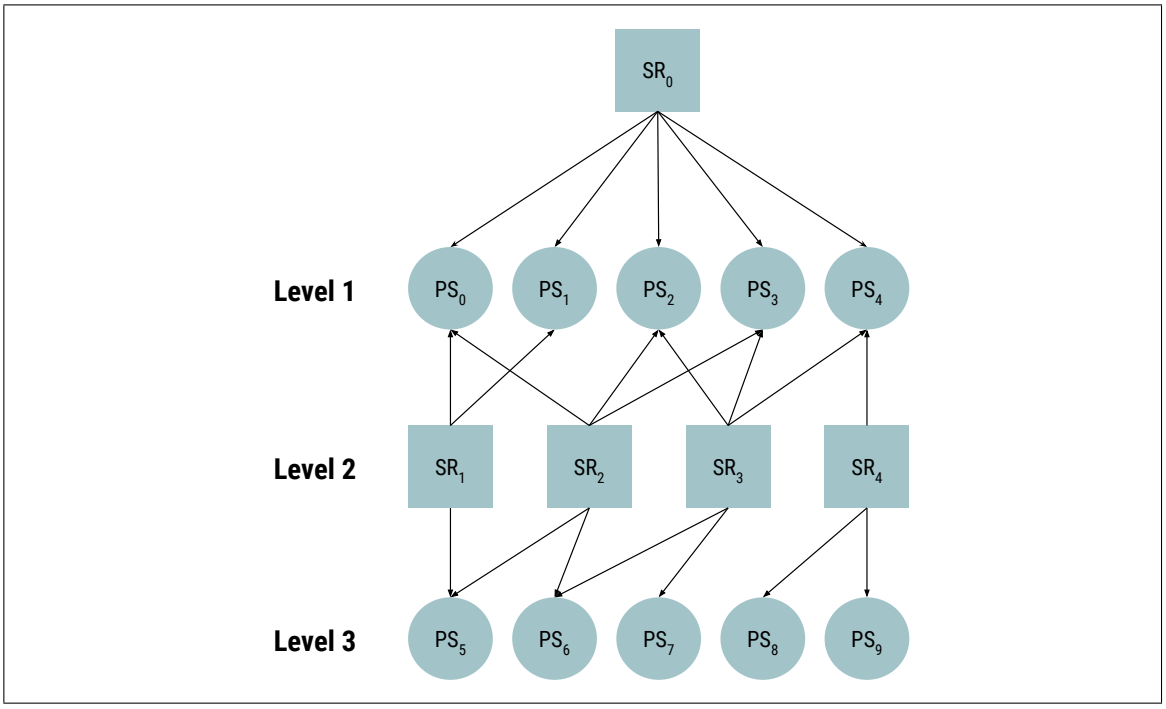

Figure 1.4. Graph-based process to create an initial evidence matrix M_0. Arrows indicate that a systematic review used in its meta-analysis that primary study

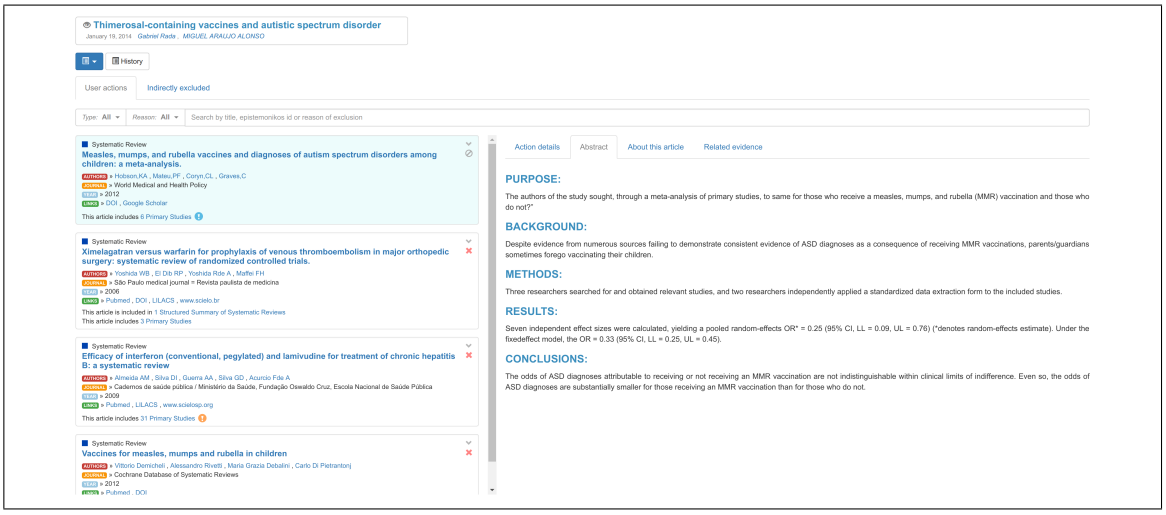

Figure 1.5. Epistemoniko's current user interface. On the left side are shown the papers that could be relevant to the question. On the right, users can see the abstract and other metadata. They click on the paper and a pop-up allows them to check whether is relevant or not.

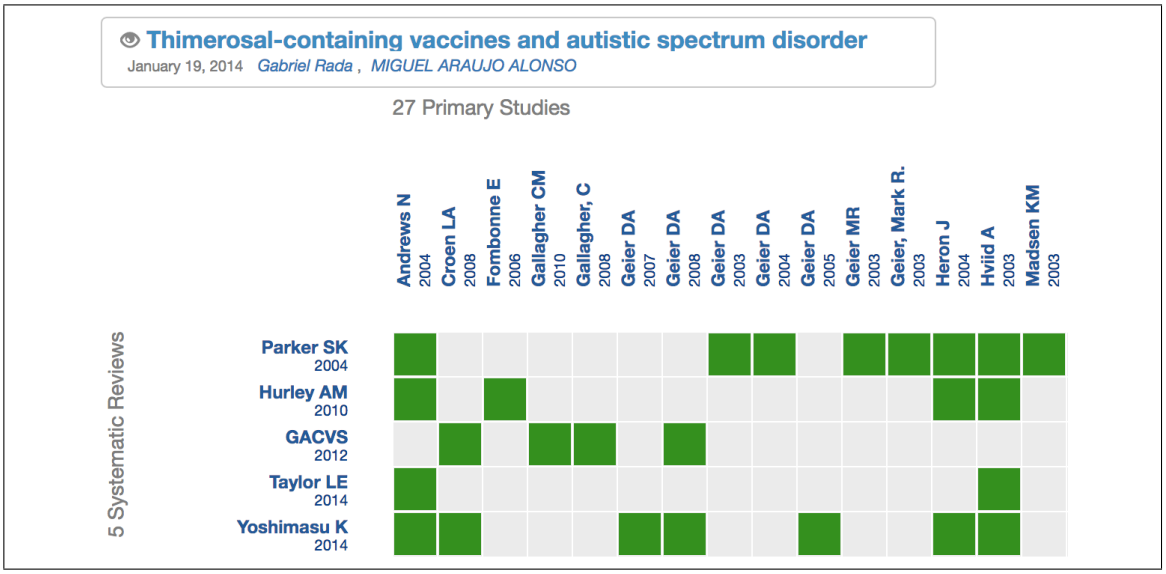

Figure 1.6. Evidence matrix for thimerosal vaccines and autism.

1.3. The problem

Nowadays, the process of creating a final matrix M_f is iterative, since involves the manual process of curating an automatically created matrix M_0. This is very slow because it requires a large amount of manual and iterative effort from experts.

Epistemonikos relies on the citation graph to find articles, but not all the articles have their primary studies linked to them. For this reason M_0 must be modified by clinicians until getting the final evidence matrix M_f by: (i) removing papers not related to the clinical question, and (ii) adding new SR and PS strongly related to the clinical question. This process can take several months, especially (ii), since it involves manually searching and checking for papers which are not explicitly linked in the *Epistemonikos*' citation database.

The main objective of this research is to design and evaluate a new way to find documents for a medical question, that requires fewer experts and less time to achieve good results. We will evaluate a combination of automatic classification and user interfaces to

find papers related to a medical question in *Epistemonikos*' database. We tested the solution in the hardest scenario, where a document does not have links in the citation graph, and only looked for documents using used text features.

2. RELATED WORK

For clarity we will first explain the state of art related to reducing the workload of citation screening and then we will introduce research associated with controllable user interfaces for information retrieval.

2.1. Reducing the workload of citation screening

Several approaches have been proposed to reduce the workload associated with the task of document filtering for citation screening in EBHC. Automatic Classification has been explored by several authors (Cohen (2006); Yu et al. (2008); Kilicoglu et al. (2009); Kouznetsov et al. (2009); Bekhuis and Demner-Fushman (2010); Kouznetsov and Japkowicz (2010); Choi et al. (2012); García Adeva et al. (2014); Bekhuis et al. (2014)). They have compared known classifiers (Naive Bayes, Support Vector Machines, K-Nearest Neighbors Bishop (2006)) with different sets of features (title, abstract, title and abstract and MESH [1] terms).

Active learning has been widely tested. Wallace, Trikalinos, et al. (2010) tested an uncertainty active learning strategy for Support Vector Machines. They asked the user to classify the instances that were closer to the hyperplane. Their results showed that this algorithm reduced the amount of documents to be screened by 40% without excluding any relevant articles. Wallace, Small, Brodley, and Trikalinos (2010) proposed two new metrics to measure the performance of Active Learning: coverage and utility. These metrics measured the usefulness of the system and number of documents that were found. Miwa et al. (2014) compared two approaches: certainty and uncertainty based active learning with several enhancements. They found that certainty approaches were better, specially in imbalanced datasets. Their LDA based enhancement was found useful, specially in complex topics. Wallace et al. (2011) presented *Meta-Cognitive Active Learning (MEAL)*, an active learning algorithm that assumes that there are experts and novices in citation screening.

[1] https://www.ncbi.nlm.nih.gov/mesh

This system allows users to classify documents as relevant, non-relevant or *too-dificult-to-classify*. Instances classified with the last label, are passed to a more expert to relabel them.

Relevance feedback was used by Jonnalagadda and Petitti (2013) to classify documents. They obtained 95% recall reducing from 6% to 30% the amount of documents that had to be screened. Data Visualization was used as a classifier in the works of Felizardo et al. (2012) and Felizardo et al. (2013). They presented a visual text mining tool that used text and the citation network as input.

To our knowledge, there are two systems that include tools to help filter documents for a systematic review. Wallace et al. (2012) presented *Abstrack* a system that uses active learning and a simple interface to classify documents. Recently, Howard et al. (2016) presented *SWIFT-Review*, a program that ranks documents according to a set of documents that had already been classified.

Most research on this topic was made in small datasets. Olorisade et al. (2016) analyze the quality of the research made in the area and states that *"More than half of the studies used a corpus size of below 1,000 documents for their experiments while corpus size for around 80% of the studies was 3,000 or fewer documents"*. Using these small datasets does not seem appropriate in this area because (i) these datasets where the output of a previous strategy search performed on databases, so the text corpus is smaller than if would be in real environment; and (ii) the number of papers is not comparable to the size of current medical databases[2]. In our research we use *Epistemonikos*' database which has nearly 400, 000 documents. For this reason, this is the first research in the area made on a real, big and noisy database.

[2]https://www.nlm.nih.gov/bsd/index_stats_comp.html

2.2. Controllable User interfaces for information filtering

Ideally, we would like to create an automatic system to solve the problem of finding relevant research articles for answering clinical questions. However, domain experts usually like to have more control than non-experts on systems supported by intelligent algorithms (Knijnenburg & Willemsen, 2011). This is the case for physicians looking for papers to answer a medical question.

Controllable interfaces have shown to increase satisfaction in recommender systems, because they increase transparency and trust (Hijikata et al., 2012; Knijnenburg et al., 2012; Bostandjiev et al., 2012). This type of interfaces, also increase user engagement and leads to better user experience (Parra & Brusilovsky, 2015).

Following this inspiration, some authors have created interfaces that support information filtering. di Sciascio et al. (2016) proposed a interface to present search results. In this interface, all results can be ranked according to words that appear on the document. Users can assign a weight to each word they select and the documents will be re-ranked accordingly. Peltonen et al. (2017) presented an interface to support relevance feedback that has a visualization of topics and keywords. Users can give positive feedback as well as negative feedback. They proved that adding negative was beneficial for complex tasks. Beltran et al. (2017) presented an interface with swipe gestures that allows users to classify documents in two groups. This intelligent system creates bins in each group that allow users to justify their classification without needing to write anything.

Compared to these works, we provide the first controllable and transparent information filtering system for EBHC, inspired by controllable recommender system interfaces. This is the first system that works with this type of data in real conditions and involving users in the process.

3. OBJECTIVES

In this area, it is essential to keep a "human in the loop" in the process, since physicians require control and transparency while filtering documents to answer a clinical question. Previous works attempted to solve this issue using automatically or with semi-supervised approaches, and even though they showed promising results, these approaches are still not widely used in practice.

The general objective of this work is to present a novel way to reducing the amount of time and people needed to find the research articles which answer a clinical question.

3.1. Hypothesis

(i) **Does a document visualization affect the task's outcome?**
A data visualization can help users to understand recommendations made by the system. We want to test whether a visualization can improve recall and/or user satisfaction with the task in EBHC.

(ii) **Does a relevance feedback algorithm affect the task's outcome?**
We want to test whether using a relevance feedback algorithm can have the same results that automatic classification has shown.

(iii) **Are there other factors that can affect the task's outcome?**
We want to analyze if there are other factors, like the search's topic, reading skills in English or previous user knowledge, that could affect this task outcome.

3.2. Contribution

In this work, we introduce *EpistAid*, a system with an intelligent user interface which support physicians in the process of finding documents for a medical question. Then, we

contribute to EBHC and to the area of intelligent user interfaces by: (a) integrating dimensionality reduction and leveraging relevance feedback for assisting incremental document classification in EBHC, and designing and implementing an interactive user interface which integrates the aforementioned methods to reduce the effort required to finish this important task of health care.

4. SOLUTION

We call our solution *EpistAid*. We propose a series of methods, which combined with an interactive user interface, aim at reducing the time to obtain the list of documents related to a new medical question from a related systematic review.

Our solution encompasses a user interface and algorithms that can assist physicians during the process of removing and adding documents related to a clinical question they want to answer. We present *EpistAi*d in three parts: (i) User interface, which describes the layout and visual components, (ii) Interactions, where we justify and describe our design based on Schneidermann's visual Information-Seeking mantra (Shneiderman, 1996), and (iii) Algorithms, which support the intelligence behind the filtering process.

4.1. User interface

EpistAid is a web application. The user interface was developed using D3.js (Bostock et al., 2011) and Bootstrap (Otto & Thornton, 2016). The GUI layout, shown in figure 4.1, has 7 parts described as follows:

- **(A) Navigation Bar.** The navbar shows ther search's title and the control buttons: search for more documents, continue latter, finish the search and get help. It also shows how much time the user has left to do the search. The number of papers the user has to give feedback for is shown in a bagde inside the button to search for more documents.
- **(B) Suggested Documents' Bin.** The system will look for documents related to the topic and will display them in this bin. Each document is represented by a rectangle, with a fixed width. The title is cut when its length is bigger the width of the rectangle. When clicking the document, its details will appear on **(F)** and its figure in the **(E)** will be bigger. The documents are sorted from left to right

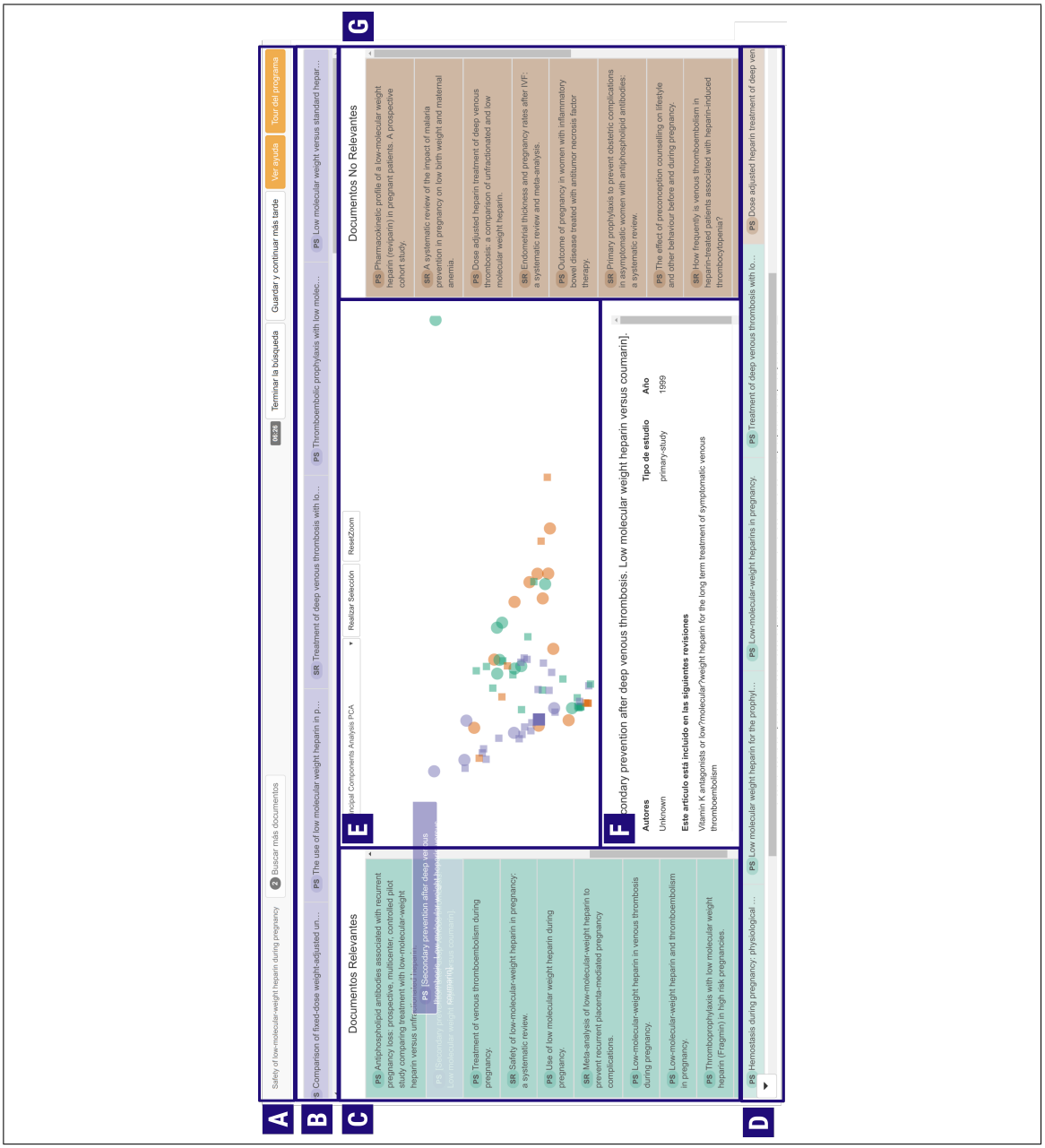

Figure 4.1. *EpistAid* layout overview. (A) Navigation Bar section allows to get more documents. (B) Suggested Documents' Bin section contains the documents suggested by the algorithm. (C) and (D) contain all documents classified by the user. (E) Documents Visualization area shows documents as figures in a 2D chart. (D) section shows all documents that have been classified in order. (F) shows the metadata of a selected document.

according to the *relevance model* of the document set, a concept we explain in the next section 4.3.

- **(C) Relevant Documents' Bin and (G) Non-Relevant Documents' Bin.** These bins have the documents that have already been classified by the user as relevant o non-relevant. Documents inside the bins don't have any particular order and users can simply drop a document on them.
- **(E) Documents Visualization.** This area shows the documents as figures in a 2D chart. Its purpose is to provide an overview of the documents that are in any of the bins (suggested, relevant and non-relevant), and to let the user explore the content based on proximity among documents. Since we represent this set of documents as a document-term matrix (DTM) using a vector space model (Manning et al., 2008), we perform dimensionality reduction over this DTM to represent each document with a low-rank vector of two dimensions. We chose 5 different dimensionality reduction algorithms: Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA), Multidimensional Scaling (MDS) (Engel et al., 2012) and the recent t-distributed Stochastic Neighbor Embedding or t-SNE (Maaten & Hinton, 2008). Users can choose the type of dimensionality reduction they prefer in order to eventually visualize the documents in the two dimensional (2D) chart shown as **(E)**.

  In the 2D chart, the primary studies (PS) are represented by circles and the systematic reviews (SR) by slightly larger squares. The color is used to discriminate the current status of the document: relevant, non-relevant or unknown.

  To select a set of documents, the user can draw a rectangle with custom dimensions, what we call a *brush*. The *brush* enables the user to navigate through all the documents by sub-setting them based on their positions in the 2D projections. The selected document are highlighted in all bins (relevant, non-relevant, suggested) and in the timeline **(D)**. It is also possible to zoom in and out the visualization.

- **(D) Actions Timeline.** This panel shows all the documents the user has classified. The color of each item represents the class that was given to the document and the order is chronological from left to right.
- **(F) Document Detail.** For a document selected in any of the panels, this area shows its meta-data (title, abstract, type of study, publication year and authors). Its goal is to offer the user the option to review the documents in the same way they usually do.

4.2. Interactions

Our interaction design is based on the visual Information-Seeking Mantra: overview first, zoom and filter, then details-on-demand (Shneiderman, 1996).

- **Overview first.** We implement the *overview-first* functionality with the *Documents Visualization* **(E)**, where users can see a summary of the documents from the selected evidence matrix in the 2D projection resultant from a dimensionality reduction over the term-document matrix. Users can also see both bins, relevant and non-relevant with all the documents that were classified.
- **Zoom and Filter: selecting documents.** The *brush* tool, described in the previous section allows, users to subset documents from the 2D *Documents Visualization*. These documents will be highlighted in all panels, so the user can see their titles. She can also zoom the visualization to look a small quantity of papers and select them by clicking or using the brush.
- **Details on demand.** By allowing the users to click in the documents on any of the bins or in the Actions Timeline, the systems provides additional details displayed in *Document Detail* (**F**).

4.3. Algorithms

The application backend was programmed in Python using Scikit-Learn (Pedregosa et al., 2011), Pandas (McKinney, 2010) and Scipy & Numpy (van der Walt et al., 2011). The documents were represented using Term-Frequency.

The main aspect of our "human-in-the-loop" algorithmic procedure starts with modeling the medical question as a *query*, which is updated iteratively when users provide relevance feedback. We tested two algorithms: Rocchio (Salton, 1971) and BM25 (Jones et al., 2000). We choose this type of model because it mimics the actions users currently do when they search for documents, in the sense that users look for documents, see whether they are relevant or not, reformulate the search query and so on, until they think the found all related evidence.

For both models we define a query q_i as a vector of words, as $\overrightarrow{q_i} = \{w_1, w_2, \ldots, w_n\}$ where n is the number of words in the corpus and w_j is the frequency of the word j. In our system the initial query q_0 is made from the words in the title and abstract of the Seed SR. We chose this representation because it is not database depedant.

4.3.1. Rocchio

Rocchio creates on each iteration a new query: the lower the distance of the query and the document, the better. In this way, the query is what an ideal document would look like.

This model ranks all the documents from the database based on a distance metric with q_0, predicting the top d most similar as relevant. We then recommend these documents to the users so they can confirm our predictions. Once the user provides feedback by manually classifying the recommended documents from the *EpistAid* interface, we update the query iteratively, such that in iteration n the query is:

$$\overrightarrow{q_n} = \alpha\overrightarrow{q_{n-1}} + \beta\overrightarrow{q_0} + \frac{1}{|R|}\sum_{\overrightarrow{d_r}\in R}\overrightarrow{d_r}\vec{\gamma}^T - \frac{1}{|I|}\sum_{\overrightarrow{d_i}\in I}\overrightarrow{d_i}\vec{\delta}^T \tag{4.1}$$

Table 4.1. Terms and their meaning of BM25 scoring function

	Meaning
N	Number of documents in the database
df_t	Number of documents that have term t in their representation
$\|VR\|$	Number of relevant documents
$\|VR_t\|$	Frequency of term t in relevant documents
$\|VNR\|$	Number of non relevant documents
$\|VNR_t\|$	Frequency of term t in non relevant documents
tf_{td}	Frequency of term t in document d
tf_{tq}	Frequency of term t in current query
L_d	Document d length
L_ave	Average document length in the database

Where q_{n-1} is the last computed query, R is the set of relevant documents, I is the set of non-relevant documents, q_0 is the initial query, d_i is the document term matrix (DTM) for the non relevant documents and d_r in the DTM for the relevant documents. The parameters α, β, γ and δ have values between 0 and 1.

4.3.2. BM25

This model uses a weighting scheme for words, to obtain a score for each document. It can be sensitive to the amount of words in each document, but in our case we do not have this problem because we are dealing with abstracts and titles that normally have between 100 to 500 words. It obtains the score of one document RVS_d using the following formula:

$$RVS_d = \sum_{t \in q} \log \left[\frac{(|VR_t| + 0.5)/(|VNR_t| + 0.5)}{(df_t - |VR_t| + 0.5)/(N - df_t - |VR| + |VR_t| + 0.5)} \times \frac{(k_1 + 1)tf_{td}}{k_1((1-b) + b(L_d/L_{ave})) + tf_{td}} \times \frac{(k_3 + 1)tf_{tq}}{k_3 + tf_{tq}} \right] \tag{4.2}$$

the meaning of each term appears on table 4.1.

The first term normalizes the word importance in the dataset given its frequency in the documents classified and the dataset; the second term weighs the word based on its frequency in the document and length of the document; the third term weighs the word based on the frequency in the query.

After each iteration, words can be added to the query using the score function of 4.3. We can select the i words with higher values. This function corresponds to the first term of the formula 4.2.

$$ScoreTerm_t = \frac{(|VR_t| + 0.5)/(|VNR_t| + 0.5)}{(df_t - |VR_t| + 0.5)/(N - df_t - |VR| + |VR_t| + 0.5)} \tag{4.3}$$

Reasonable values for parameters vary between 1.2 and 2 for $k1$ and $k3$ and 10 to 25 for i. Normally b is set to 0.75 (Manning et al., 2008, p 233).

4.4. System Architecture

The system has three main parts. The *Processing Module* is in charge of keeping track of users sessions and also performs relevance feedback and ranks the documents. This module serves from two data sources: a mysql database and pickle binary files. The mysql database has the documents' metadata and it is used to display the information on the user interface. The binary files have a Compressed Sparse Row CSR matrix that has the term frequency of each word in the corpus (columns) for each document (rows).

The *API* was developed using Flask [1], a microframework to develop web applications. It is called by the *Interface* via AJAX, whenever the user wants to add feedback.

[1] https://github.com/pallets/flask

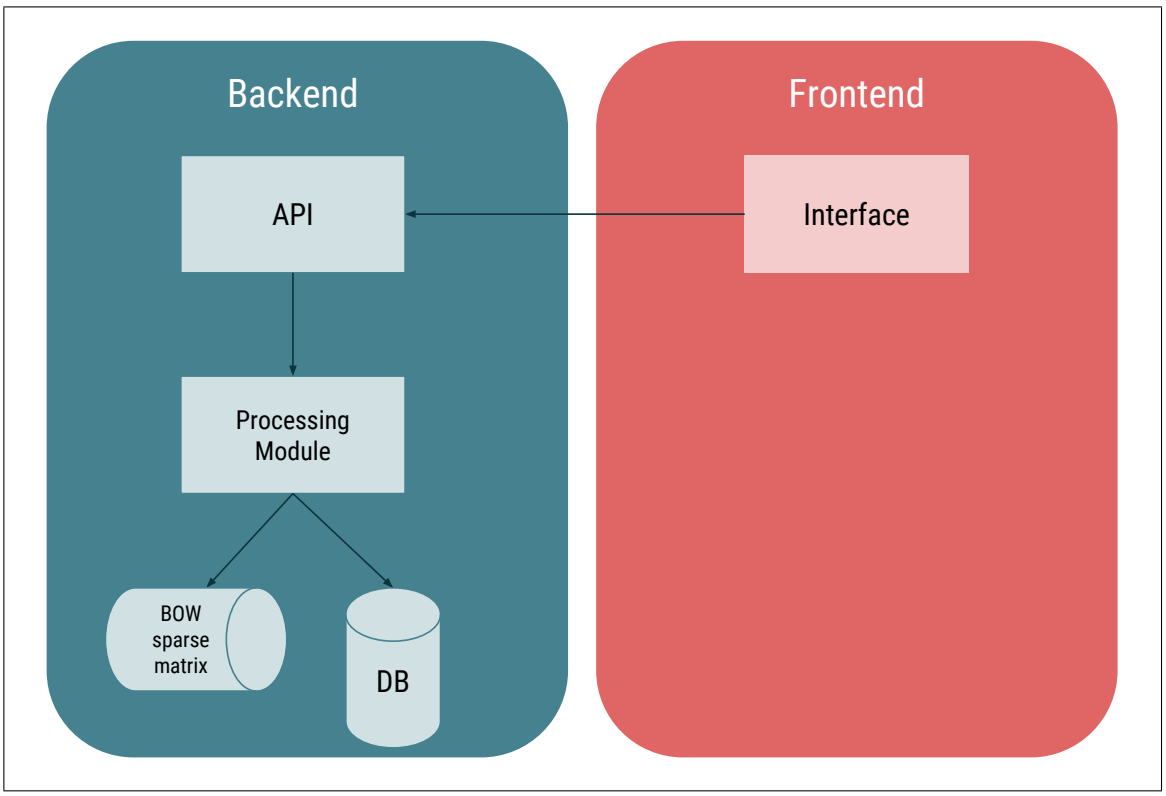

Figure 4.2. *EpistAid* architecture. The interface consumes an API that consults the processing module. The data is stored in a relational database and one precomputed file.

5. METHODOLOGY

5.1. Dataset

We used a dump of the database of *Epistemonikos* of December 2016. *Epistemonikos* collects articles from 26 online sources (Epistemonikos, 2016). The database contains around 390,000 documents of five types:

- **Broad Syntheses**: articles that aim to synthesize several systematic reviews.
- **Systematic Review**: Any article which analyze and summarise several primary studies
- **Structured Summary of Systematic Review**: corresponds to a summary of a systematic review for an audience of non-researchers.
- **Primary Study**: This term encompasses any type of study (quantitative or qualitative) where data from individuals or groups is collected.
- **Structured Summary of Primary Study**: corresponds to a summary of a primary study for an audience of non-researchers.

Table 5.1 shows the number of items per publication type in *Epistemonikos*.

5.1.1. Matrices

Currently, there are about 2,700 public evidence matrices, but only 1065 are in their final revised version. Physicians require 2-6 months to get from an initial version M_0 to a final revised M_f. It takes 1-2 days to filter the initial matrix M_0 (removing non-relevant

Table 5.1. Distribution of publication types in the database.

Type of publication	**Articles in database**
Primary Study	277, 967
Systematic Review	73,040
Overview	1, 229
Structured Summary of PS	1,351
Structured Summary of SR	37,779

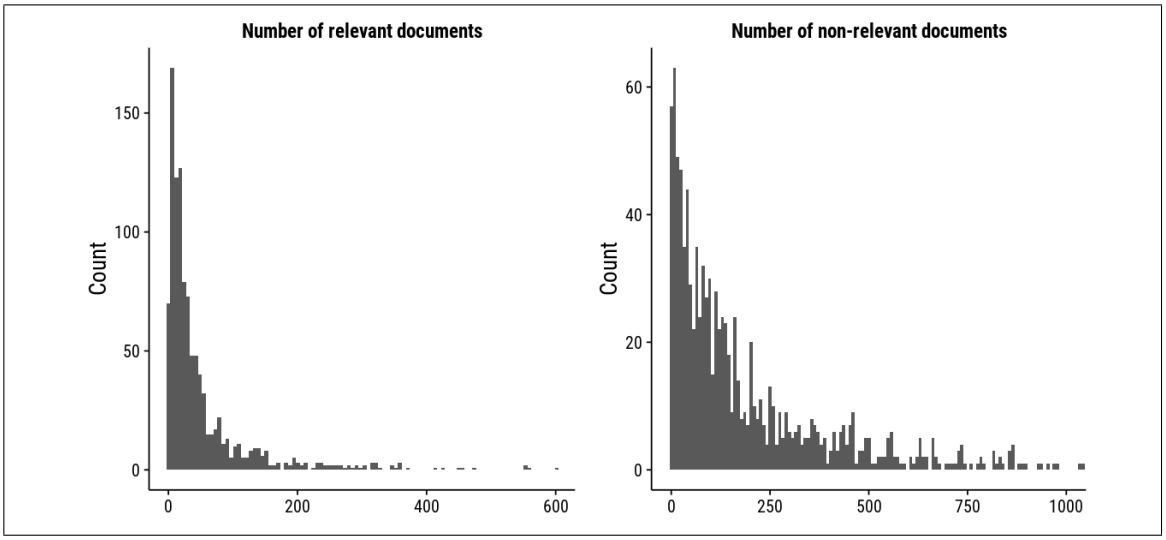

Figure 5.1. Histogram of relevant and non-relevant documents for each *Evidence Matrix*

articles), and the rest of the time is spent in finding other papers that could be related, but which were not found using the graph approach. This second stage of adding new relevant documents consists of two steps: creation of a new search strategy (list of keywords) that will be used to find other articles, and using of the new strategy in the Epistemonikos' database and other databases to collect a list of papers. They are filtered and then inserted in the graph, so that they can appear in the matrix.

Each matrix has a set of relevant documents and non-relevant documents. These labels have been applied by a doctor or a student. In the dataset of December 2016, matrices had on average 68.23 relevant documents ($SD = 185.99$) and 258.80 non relevant documents ($SD = 462.08$). The distribution count is shown in figure 5.1. Not all matrices are comparable because the variance is high. This makes the evaluation harder because not all matrices will be comparable in terms of the amount of documents that need to be found.

In our experiments, an Epistemonikos' matrix corresponds to a medical question with its relevant documents. All documents that are not listed in the matrix are considered non relevant. The seed systematic review is the initial query in the relevance feedback algorithms.

5.1.2. Preprocessing

Each document was represented by the words in its title and abstract. There are not differences between these words in the vector. If one of them was not in the database, we used only the other one. We split the text by non alphanumeric characters (spaces, numbers and symbols). Then, we removed all stopwords listed for English in the NLTK package (Bird et al., 2009). Many articles used acronyms of medical concepts that had less than three characters. The same acronym can be used in different medical domains, which means one token can have different meanings. For this reason we decided to remove any word that had less than three characters. All words were stemmed using *Porter Stemmer* from the same package. This whole process resulted in a corpus of $158,457$ words. We removed all documents that did not have words of the corpus and obtained $391,364$ documents.

5.2. Algorithm Evaluation

We did an offline evaluation, where we simulated users that gave feedback. The user we simulated was a perfect user that knew exactly what documents were relevant. This evaluation was performed in order to have a baseline on the system performance. We split the data set in training and testing with a ratio of $70/30$. We used the training data to select the best parameters of each model. The other 30% of matrices were used to evaluate performance of the chosen parameters.

To test each model, we simulated relevance feedback iterations: the relevance feedback model gave a ranked list, the *perfect user* would say what documents were relevant, to finally add that feedback to the model which generated a new ranked list from the documents that hadn't been seen yet. This process continued until the model didn't find any relevant documents. The idea behind this simulation is to evaluate what would happen if we had a perfect user, who knew exactly what documents were relevant.

For Rocchio we tested different values for α, β, γ and δ that varied from 0 to 1 with steps of 0.25. We only kept combinations that adding the values for all parameters gave a total of 1.25 or less, to be able to compute the experiments in reasonable time. Cosine similarity was used to compare the query to all the documents in the database.

In the case of BM25, we used $b = 0.75$ and varied $k1$ and $k3$ with values from 1.2 to 2.0 with steps of 0.1. We also tested the amount of words that could be added to the query after each iteration. We tested adding 15, 20 and 25 words on each iteration.

For both models, after each iteration, feedback for 30 documents was added. We found that this number had higher coverage and it was a reasonable number of items to show to a user at once.

We evaluate performance using MAP@30 in the offline evaluation. In the user study, we evaluate performance of the algorithm using accumulated recall and MAP@30

5.3. User Study

We conducted a user study to compare user performance on the relevance feedback algorithm and on the visualization. It was a 2x2 Within Subjects Factorial Design, where users were asked to find documents related to one medical question. On each experiment the user had to test a combination of an algorithm and an interface as seen in figure 5.2. The algorithms tested were BM25 and Rocchio and the interfaces were the one describe in the previous section (figure 4.1) and the interface without the visualization (figure 5.3).For each experiment, users answered different medical questions. Question A corresponds to *Safety of low-molecular-weight heparin during pregnancy* and question B to *Echinacea for common cold*. These questions are explained in table 5.2. Users had 10 days to perform both searches and had to spend a minimum of 15 and a maximum of 30 minutes per search on the interface.

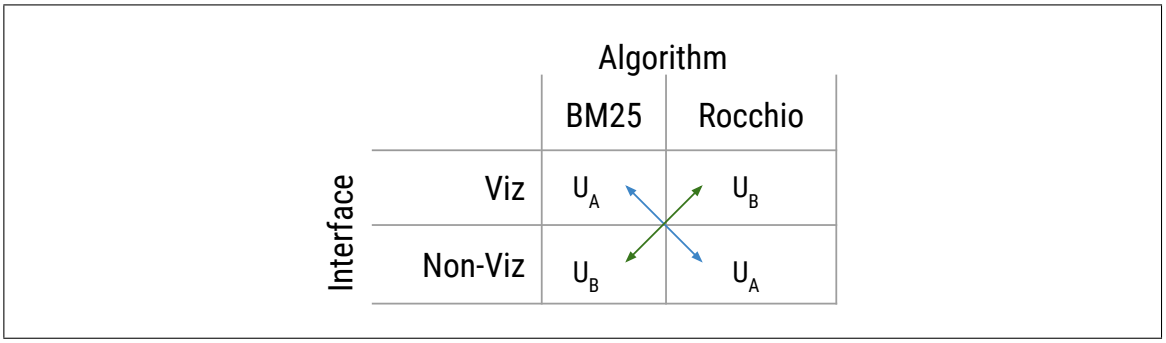

Figure 5.2. Study design: U_a will do one experiment combining BM25 with visualization and the other combining Rocchio without visualization.

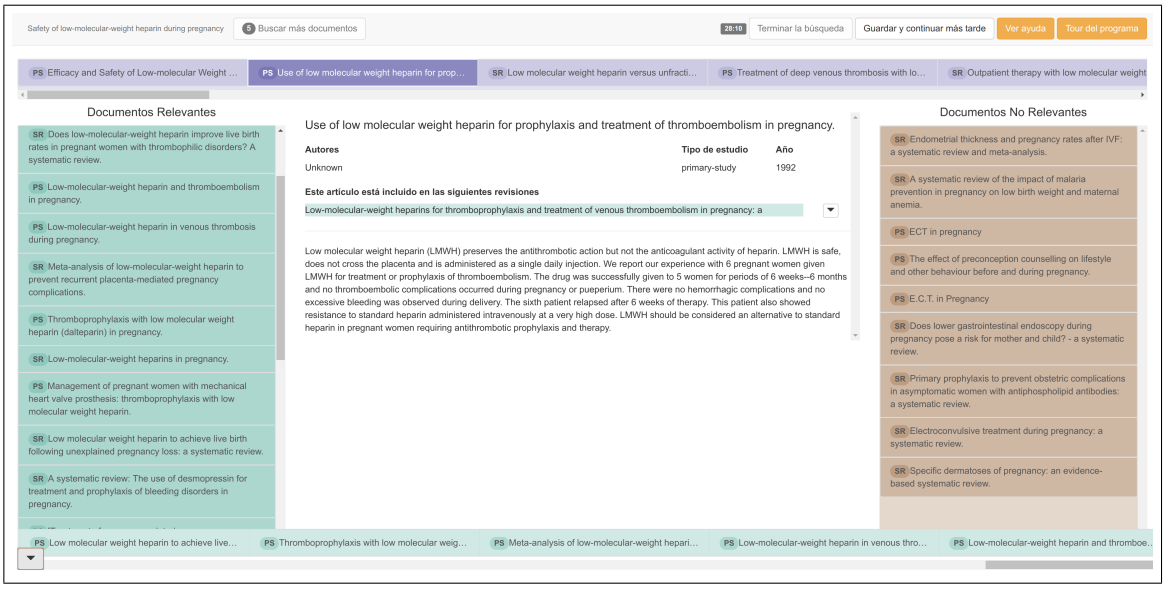

Figure 5.3. *EpistAid* layout overview without visualization.

Before starting the study, users had to fill a consent (appendix C) form and answer a survey where they were asked questions related to their previous knowledge on the questions' topics, their experience working on finding documents to answer a medical question, reading research in English and with data visualizations. The survey questions are displayed in appendix A. After they performed the search, they answered a survey with questions related to their satisfaction with the system (appendix B). We measured cognitive effort using the NASA-TLX survey. Users were also asked to explain the process they followed to classify a document.

Table 5.2. Explanation of medical questions used in the user study

Question	**Specialty**	**Context**
Safety of Low-molecular-weight heparin during pregnancy	Internal Medicine	Low-molecular-weight heparin is anticoagulant medication. This question aims to find if it is safe to use it during pregnancy, because of the possible risk to the baby.
Echinacea for common cold	General Medicine	Echineacea is a flower. There are some web sites that say that it is usefull for common cold, but there is no evidence for that.

The system tracked every click and scroll users did on interesting objects in the interface. We defined three types of actions: *explore*, *lookup* and *click*. All scrolls the user did on any of the bins were saved as *explore* actions. Scrolling on document metadata was considered a *lookup* action. Every click on a document in any of the bins (relevant, non-relevant o suggested) was considered a *click* action. In the case of the visualization, other *explore* actions were doing zoom or brush on the visualization area. Clicks on figures in the visualization were considered *click* actions.

Every time the user scrolled, the system would keep track of the scrolled item. If the user scrolled that same item again in less than 5 seconds, that scroll wouldn't be considered. This way we avoided to keep track of random scrolls.

To evaluate the algorithm, we computed precision and recall of relevant documents at:

- *Seen documents*. Taking into account all documents the user saw during the session, and their state (relevant, non-relevant, unknown) at the end of the task.
- *Ground truth*. Taking into account all documents the user saw during the session and the documents that were relevant to the question but were never suggested by the system. Like at *Session level* the state (relevant, non-relevant, unknown) was taken into account at the end of the task.

We used these two ways to evaluate because with the first we measure the effectiveness of the user during the task. With the second we measure the effectiveness of the system.

6. RESULTS AND DISCUSSION

6.1. Offline Evaluation of Algorithm

Both models had the same behavior. In the first queries their performance was better at the begining, and then it decreased. The main difference between the models is variance. Rocchio has larger variance in recall and BM25 tends to have similar performance for all queries. The results for both models in detail can be downloaded from `https://goo.gl/we1Ywg`.

6.1.1. Rocchio

Rocchio has an average MAP@30 of 0.37 (table 6.1) across all different parameters combinations at query 0.

We grouped results by each parameter and compared their Mean Average Precision. We will show only the values of MAP at 30th ranked position, because 30 documents is an amount that can be easily screened by a user. The results are shown in figures 6.2 and

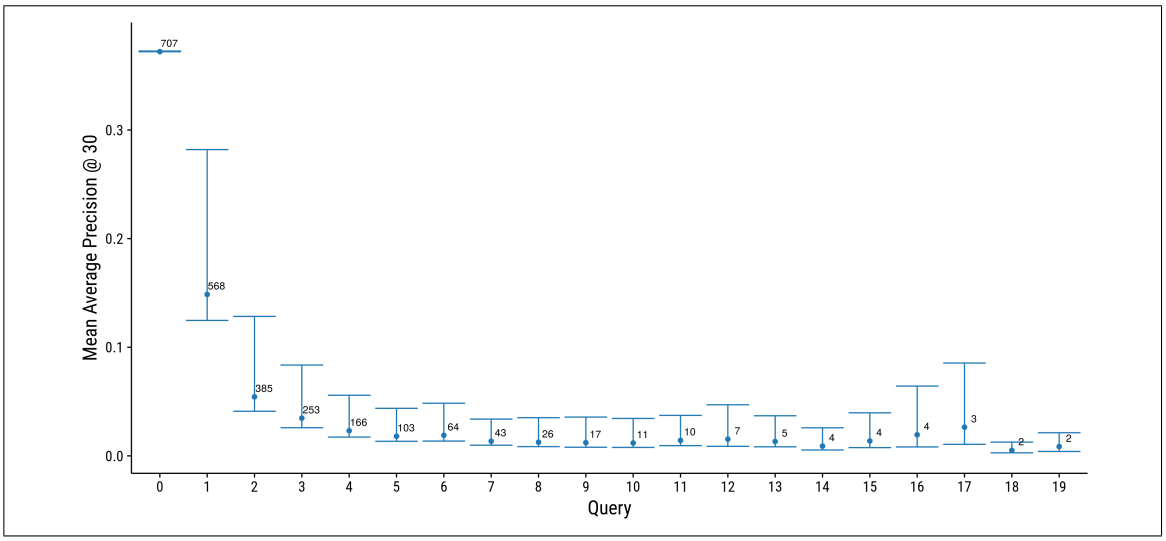

Figure 6.1. MAP@30 for Rocchio at different queries. For each query, the model looks for documents that it has not show to the user, so the model has to look for less papers.

Table 6.1. MAP@30 for Rocchio at different queries. For each query, have added the feedback of previous queries, and so the model has to look for less papers.

	Query	Mean	Std. Error	Std. Deviation	Coverage
1	0	0.37	0.00	0.00	707.00
2	1	0.15	0.02	0.13	568.00
3	2	0.05	0.01	0.07	385.00
4	3	0.03	0.01	0.05	253.00
5	4	0.02	0.01	0.03	166.00
6	5	0.02	0.00	0.03	103.00
7	6	0.02	0.01	0.03	64.00
8	7	0.01	0.00	0.02	43.00
9	8	0.01	0.00	0.02	26.00
10	9	0.01	0.00	0.02	17.00
11	10	0.01	0.00	0.02	11.00
12	11	0.01	0.00	0.02	10.00
13	12	0.02	0.01	0.03	7.00
14	13	0.01	0.00	0.02	5.00
15	14	0.01	0.00	0.02	4.00
16	15	0.01	0.01	0.03	4.00
17	16	0.02	0.01	0.04	4.00
18	17	0.03	0.02	0.06	3.00
19	18	0.01	0.00	0.01	2.00
20	19	0.01	0.00	0.01	2.00

6.3. The parameter α, which weighs the previous query becomes less important in higher queries. This is natural because the model has more information from the user, so it does not need to rely on previous information. At the first query, it does not matter which value α takes, because there is not a previous query. β, which weighs the importance of the initial query, behaves similar to α: it becomes less important in higher queries.

It is interesting to see that the latter the query, α and β become less important. It is clear that if the first query was the perfect query to find all the documents, the values of β and MAP would be positively correlated; and if there was no concept drift, that is, a change in objectives, then α would also be positively correlated to MAP. Since β and α

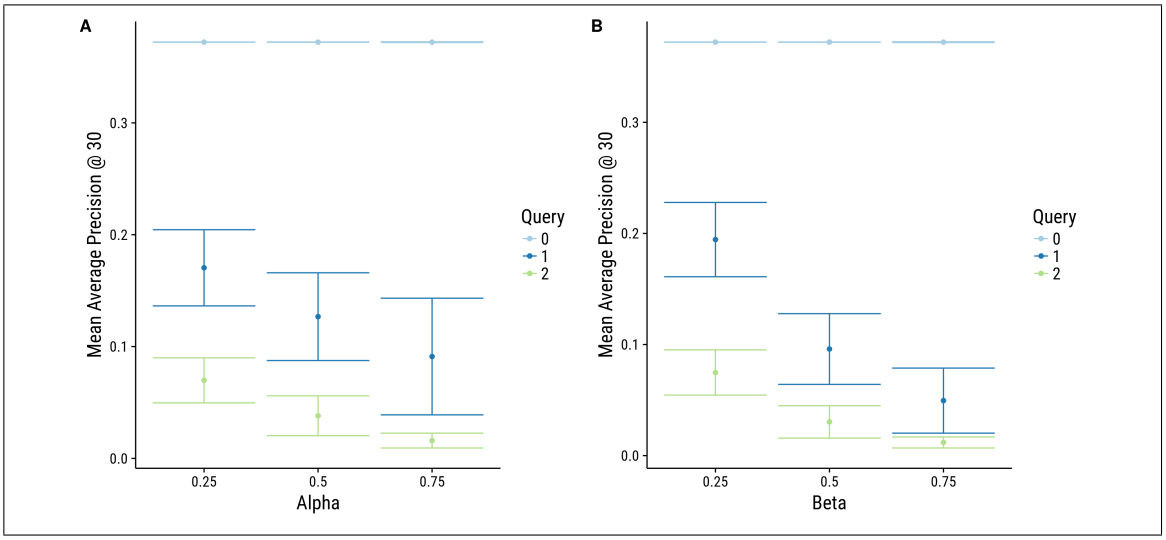

Figure 6.2. **A**: MAP@30 for Rocchio, parameter α. **B**: MAP@30 for Rocchio, parameter β.

are negatively correlated with MAP, we can assume that there is the model can deal with concept drift.

The γ parameter, that weighs the importance of relevant document has an opposite behavior to α and β: it becomes more important at higher queries. δ, that weighs the non relevant documents, has very high variance. When it is bigger than zero, that is, it subtracts to the query words that appear on non relevant documents, its variance is very high. With these results it cannot be state whether it is beneficial to MAP or not.

From these results we can conclude that previous knowledge, weighed by α and δ, are not as good as new knowledge from classified documents at making a better query. This is true specially for γ, which weighs positive feedback. Negative feedback, that translates in removing words from the query, is not always beneficial. This could happen because normally documents that are being recommended to the user, specially in lower queries, have similar vocabulary. Removing these words would mean removing relevant words.

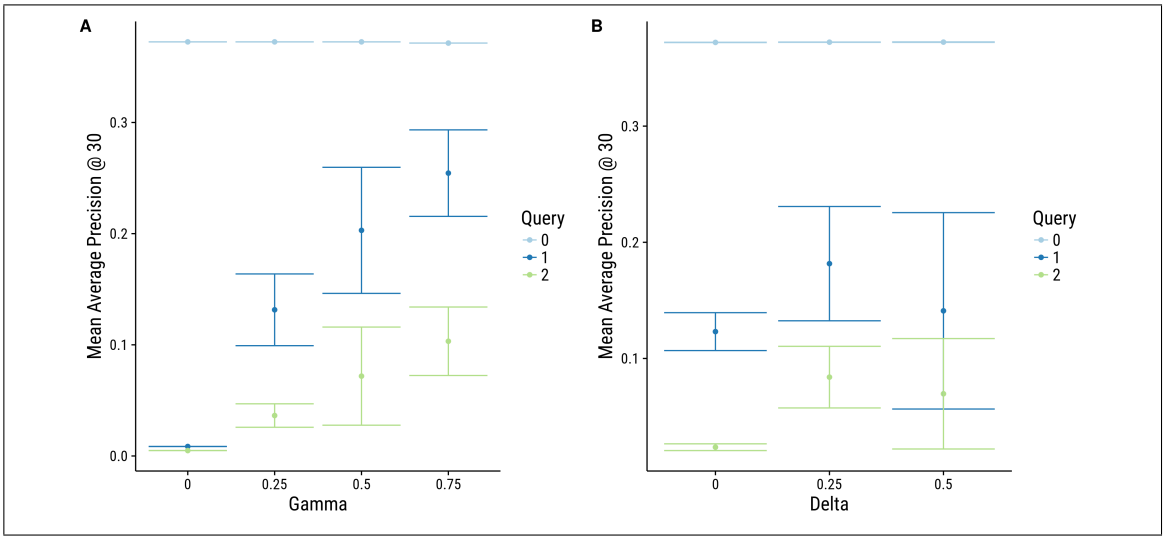

Figure 6.3. **A**: MAP@30 for Rocchio, parameter γ. **B**: MAP@30 for Rocchio, parameter δ.

6.1.1.1. Testing

We picked one the best models ($\alpha = \beta = \delta = 0.25, \gamma = 0.5$) and executed relevance feedback with the test matrices. These are the same parameters used in the user study. We found that MAP had the same behavior of the train set, this means that the model does seem to be over-fitted. This result means that the model did not learn the specific documents that were relevant for each question, but it learnt how to decide what was relevant.

We tried adding different amounts of feedback in each iteration. Specifically we did experiments in which we only added feedback for 10 and 20 documents, not 30 like we did to find the best parameters. We can see in 6.10 that adding less feedback on each iteration increases average precision, but it diminishes recall (figure 6.11). Since in this area, recall is more important than precision (O Mara-Eves et al., 2015), we decided to show 30 documents in the user study.

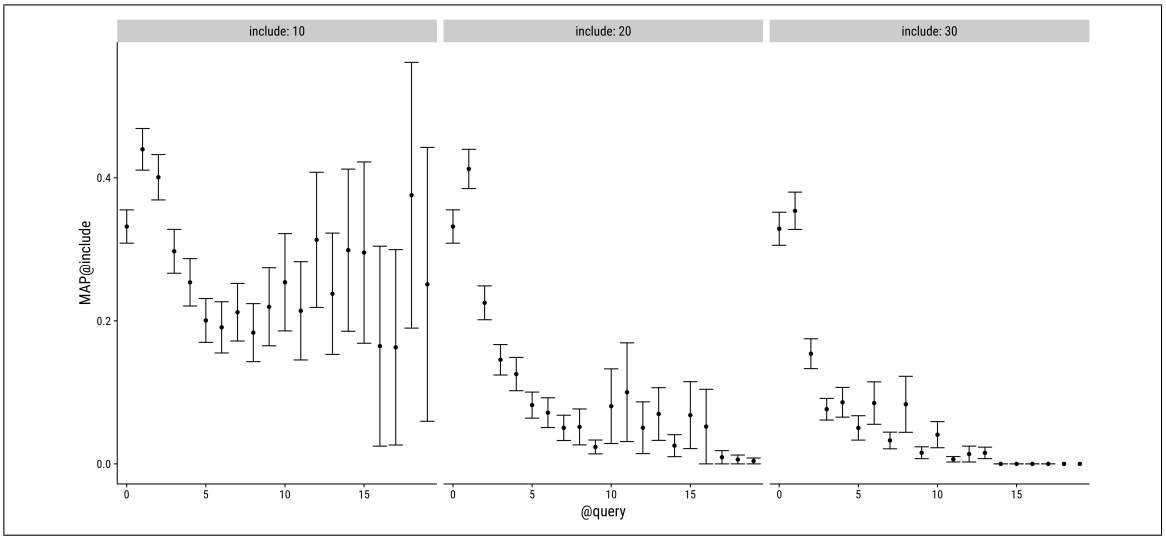

Figure 6.4. MAP@include for Rocchio in test set where *include* is the amount of documents the algorithm gave feedback for in each iteration.

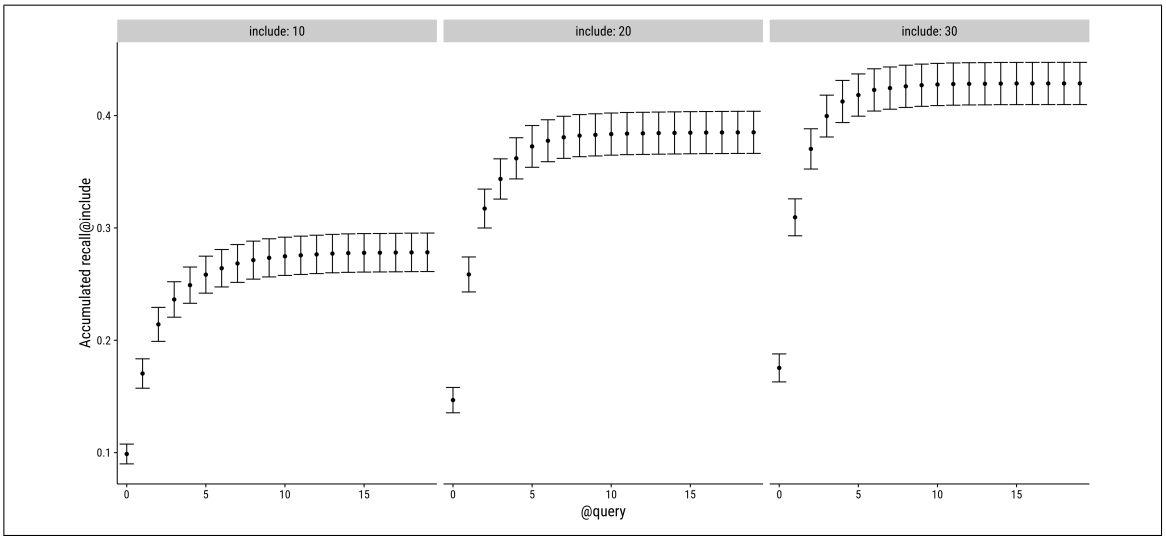

Figure 6.5. Accumulated recall@include for Rocchio in test set where *include* is the amount of documents the algorithm gave feedback for in each iteration.

6.1.2. BM25

BM25 has an average MAP@30 of 0.35 at query 0. Its coverage is less than Rocchio, which means that it was able to find relevant documents for less medical questions than Rocchio.

Table 6.2. Mean Average Precision at 30 statistics for BM25 at different queries

	Query	Mean	Std. Error	Std. Deviation	Coverage
1	0	0.35	0.00	0.01	708.00
2	1	0.20	0.00	0.01	547.00
3	2	0.06	0.00	0.01	355.00
4	3	0.03	0.00	0.01	211.00
5	4	0.01	0.00	0.00	138.00
6	5	0.01	0.00	0.01	91.00
7	6	0.01	0.00	0.01	55.00
8	7	0.01	0.00	0.01	33.00
9	8	0.01	0.00	0.01	20.00
10	9	0.01	0.00	0.02	13.00
11	10	0.00	0.00	0.00	9.00
12	11	0.00	0.00	0.00	6.00
13	12	0.00	0.00	0.00	3.00
14	13	0.01	0.00	0.05	2.00
15	14	0.02	0.01	0.06	1.00
16	15	0.03	0.01	0.07	1.00
17	16	0.02	0.01	0.04	1.00
18	17	0.01	0.00	0.01	1.00
19	18	0.00	0.00	0.00	1.00
20	19	0.00	0.00	0.00	1.00

Like we did for Rocchio, we grouped results by each parameter on its own and compared their Mean Average Precision. We will show only the values of MAP at 30^{th} ranked position, because 30 documents is an amount that can be easily screened by a user.

The behaviors of $k1$ and $k3$ have the same tendency (figures 6.7 and 6.8). For MAP it is better to have higher values in the first query, but then it is better to have lower values. This means that the characteristics of the query become less and less important ($k3$) and the same happens to the features of the document that is being scored ($k1$).

This last result is somehow similar to the results we had with Rocchio. BM25 has three terms in the score function (4.2): the modified TF-IDF, the document component, weighed by $k1$, and the query component, weighed by $k3$. Since terms weighed by $k1$ and $k3$ become less important, we can only assume that the TF-IDF component is the

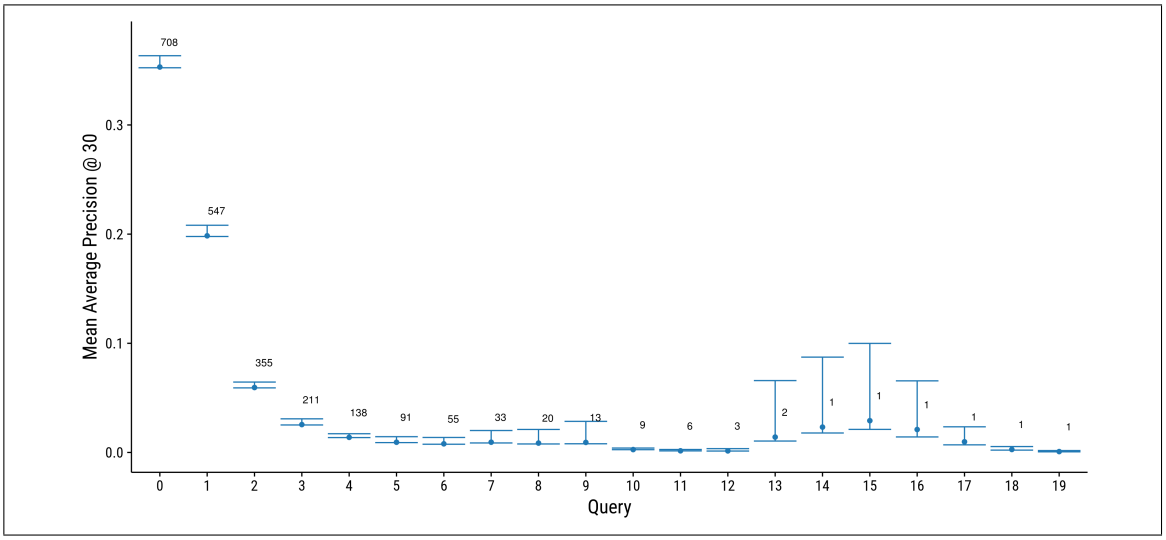

Figure 6.6. MAP@30 for BM25 at different queries. For each query, have added the feedback of previous queries, and so the model has to look for less papers.

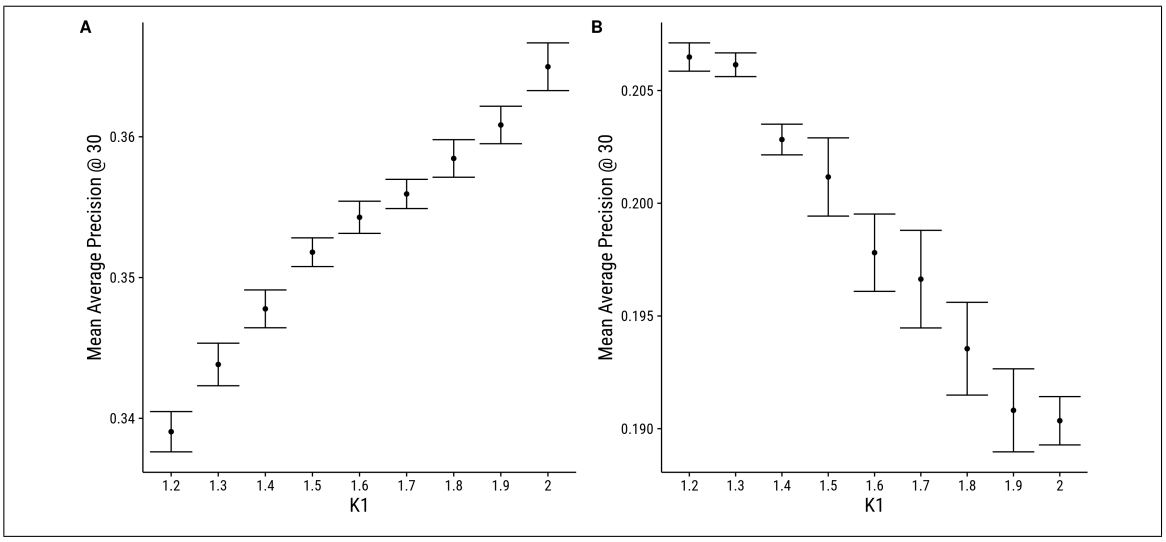

Figure 6.7. MAP@30 for BM25 for parameter $k1$. Figure A is the graph for query 0 and figure B is the graph for query 1.

most important term of the function. This term depends only on the documents that have been classified, so just like in Rocchio we have that the words in documents that had been classified are more important than other components.

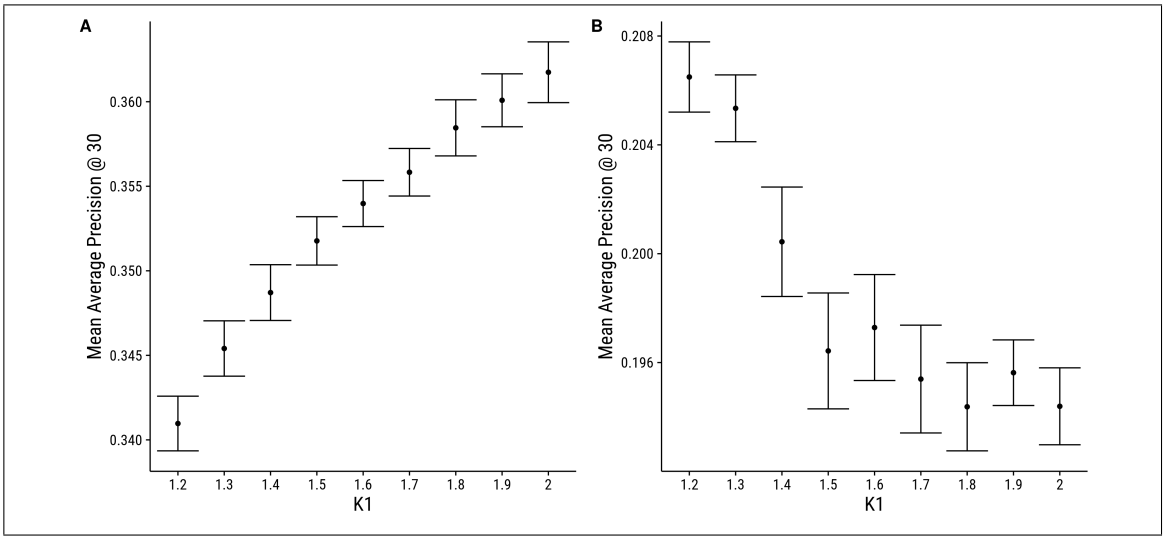

Figure 6.8. MAP@30 for BM25 for parameter $k3$. Figure A is the graph for query 0 and figure B is the graph for query 1.

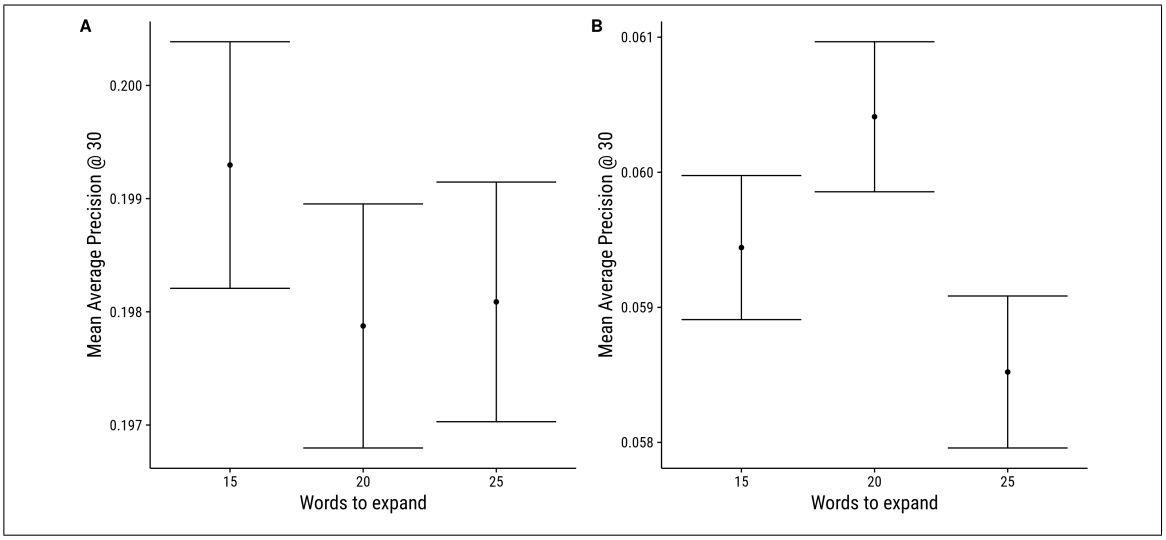

Figure 6.9. MAP@30 for BM25, parameter i. Figure A is the graph for query 1 and figure B is the graph for query 2.

The amount to words used to expand the query after each iteration does not seem to have a big effect in performance. As shown in figure 6.9, there are differences in performance but they are very small of less than 0.001.

6.1.2.1. Testing

We picked one the best models ($k1 = 1.7, k2 = 1.2, i = 25$) and executed relevance feedback with the test matrices. This values we also used in the user study. We found that MAP had the same behavior of the train set, so the model does seem to be over-fitted.

Just like we did we Rocchio, we tried adding feedback for 10 and 20 documents. The results are similar to Rocchio, MAP increases when users had less feedback, but recall diminished.

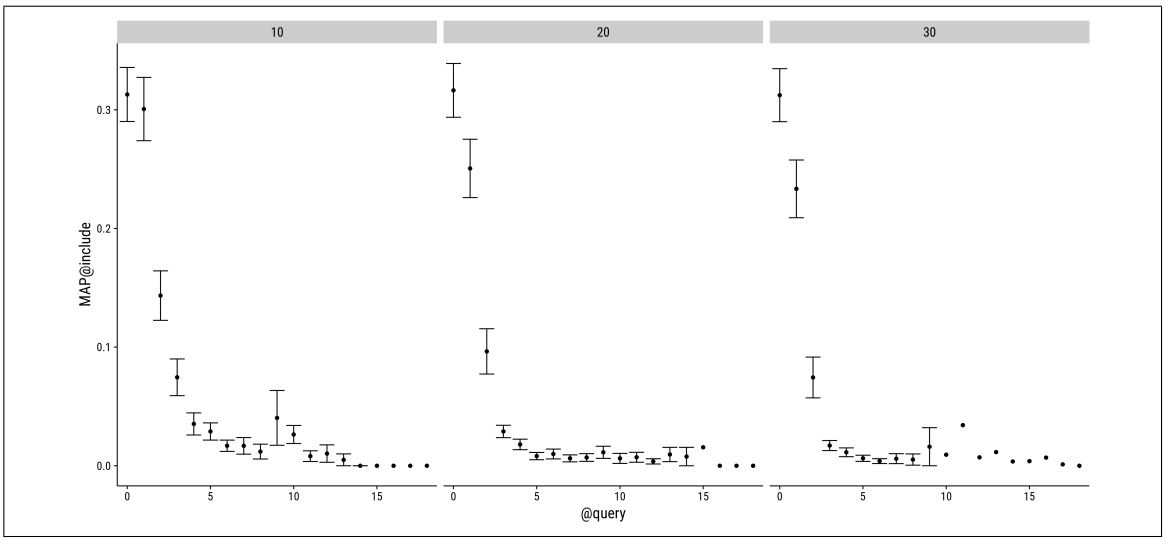

Figure 6.10. MAP@include for BM25 in test set where *include* is the amount of documents the algorithm gave feedback for in each iteration.

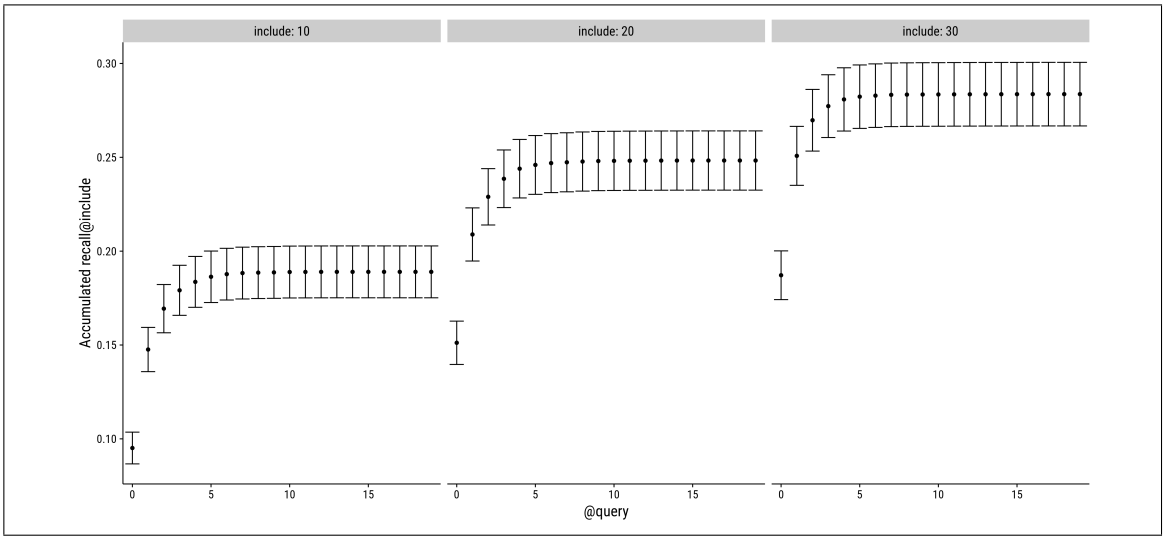

Figure 6.11. Accumulated recall@include for BM25 in test set where *include* is the amount of documents the algorithm gave feedback for in each iteration.

6.2. User Study

We conducted the user study in June, 2017. We had 22 users who completed the experiment.

6.2.1. Users

Most users were students ($N = 19$) and had not participated in the creation of a systematic review ($N = 19$). More than half of them had created more than one evidence matrix in Epistemonikos ($N = 16$). Only one had never used Epistemonikos' system before.

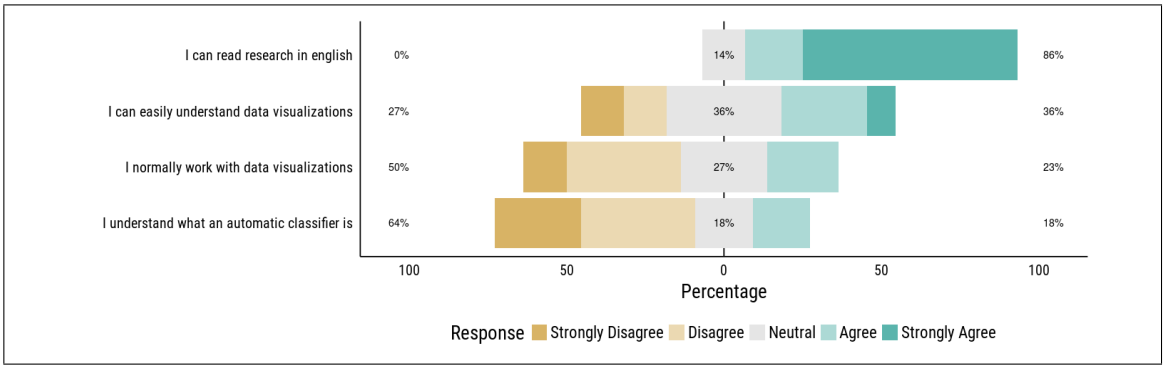

Figure 6.12. Answers pre survey

Almost all users were confident with their skills at reading research in English. At least half on them do not usually work with data visualization or cannot easily understand them.

6.2.2. Engagement and interaction statistics

. In order to measure user interaction and engagement between the four conditions, we compared three metrics: average number of queries issued per session, average number of papers classified per session, average number of papers classified per list, and time spent on the interface. Results are shown in table 6.3. We observe that in terms of time, users spent more on the conditions with BM25 ($M = 1443.91$ and $M = 1280.1$) than on the conditions with Rocchio ($M = 1115.7$ and $M = 1267.67$). In terms of queries issued, people also issued more queries in average in the BM25 conditions ($M = 5.08$ and $M = 4.9$) compared to Rocchio ($M = 4.45$ and $M = 4.41$). However, in terms of number of papers classified, list-wise and session-wise, people were more productive in the Rocchio condition with visualization ($M = 116.3$ session-wise and $M = 26.33$ list-wise). The condition with the fewest interactions was Rocchio without 2D document visualization ($M = 88.09$ session-wise and $M = 16.08$ list-wise). Another interesting metrics is that in most experiments (91%), users reported they used at least title or abstract to classify the articles.

Table 6.3. Interaction statistics in each condition studied.

Model	Interface	Time (in seconds)	Documents classified by query	Documents classified by session	Number of queries
BM25	Non-Viz	1443.92 ± 107.23	21.23 ± 1.30	107.92 ± 17.66	5.08 ± 1.18
BM25	Viz	1280.10 ± 117.17	20.44 ± 1.51	100.36 ± 16.73	4.91 ± 0.69
Rocchio	Non-Viz	1115.70 ± 80.27	19.78 ± 1.57	88.09 ± 11.98	4.45 ± 0.76
Rocchio	Viz	1267.67 ± 87.67	26.34 ± 0.96	116.33 ± 22.64	4.42 ± 0.91

6.2.3. Performance

Table 6.4 presents the results of recall, precision and F-1 score (Manning et al., 2008) at the end of the session, averaged per user. These metrics are calculated based on what the system presented to the users during the session. We also calculate recall considering the actual items in the ground truth (112 relevant documents for evidence matrix A and 54 relevant documents for evidence matrix B). Considering all the metrics, the best combination of interface and algorithm was the use of 2D document visualization with the Rocchio relevance feedback algorithm, since it has the best F-1 score considering the items seen during the session ($M = 0.7$) and the best recall with respect to the ground truth, the evidence matrices ($M = 0.23$). The interface that seemed to have the worst general performance with BM25 without visualization, specially in terms of precision, and in terms of recall considering the ground truth

Table 6.4. Average recall, precision and F-1 score (per user) considering only documents seen by users in the session, and recall considering all documents in the ground truth.

			Seen documents			Ground Truth
Model	**Interface**	**N**	**Recall**	**Precision**	**F-1 score**	**Recall**
BM25	Non-Viz	12	0.66 ± 0.08	0.52 ± 0.06	0.58 ± 0.07	0.20 ± 0.04
BM25	Viz	11	0.71 ± 0.06	0.64 ± 0.04	0.64 ± 0.03	0.18 ± 0.02
Rocchio	Non-Viz	11	0.65 ± 0.08	0.73 ± 0.02	0.65 ± 0.05	0.21 ± 0.04
Rocchio	Viz	12	0.77 ± 0.06	0.67 ± 0.01	0.70 ± 0.03	0.23 ± 0.03

6.2.4. Perception of Effort

We used the NASA TLX to measure the perception of effort, results are presented in table 6.5. The most clear pattern observed in the data is that BM25 without document visualization (Non-Viz) was perceived as requiring larger effort ($M = 48.92$), frustration ($M = 46.22$), mental demand ($M = 47.75$), and physical demand ($M = 32.83$) than the other conditions. People perceived they were performing better with Rocchio without visualization ($M = 72.5$) than with the other conditions, and this condition also required the smallest temporal demand ($M = 21.10$).

Table 6.5. NASA-TLX results grouped by RL algorithm and interface.

	BM25		**Rocchio**	
Variable	**Non-Viz**	**Viz**	**Non-Viz**	**Viz**
Effort	48.92 ± 6.83	31.80 ± 6.03	27.60 ± 5.06	31.67 ± 5.12
Frustration	46.33 ± 6.22	32.80 ± 6.41	18.00 ± 5.06	27.17 ± 6.40
Mental Demand	47.75 ± 5.15	36.60 ± 9.06	28.10 ± 6.38	41.75 ± 5.28
Performance	55.25 ± 5.45	57.30 ± 9.93	72.50 ± 6.33	63.67 ± 6.97
Physical Demand	32.83 ± 6.27	17.30 ± 7.83	18.40 ± 6.35	31.33 ± 7.93
Temporal Demand	30.25 ± 5.33	25.70 ± 8.29	21.10 ± 4.47	30.75 ± 5.61

6.2.5. Interface Interaction

Users interactions consisted mainly on clicking on documents and looking up for their detailed metadata (abstract and others). Users almost didn't check documents they had already classified (figure 6.13). On average they clicked on 0.56 documents they had already classify as non relevant and an average of 3.63 on documents they had classified as relevant. The numbers of regrets –when a user changes it previous classification– was very low ($M = 0.79$, $SE = 1.04$).

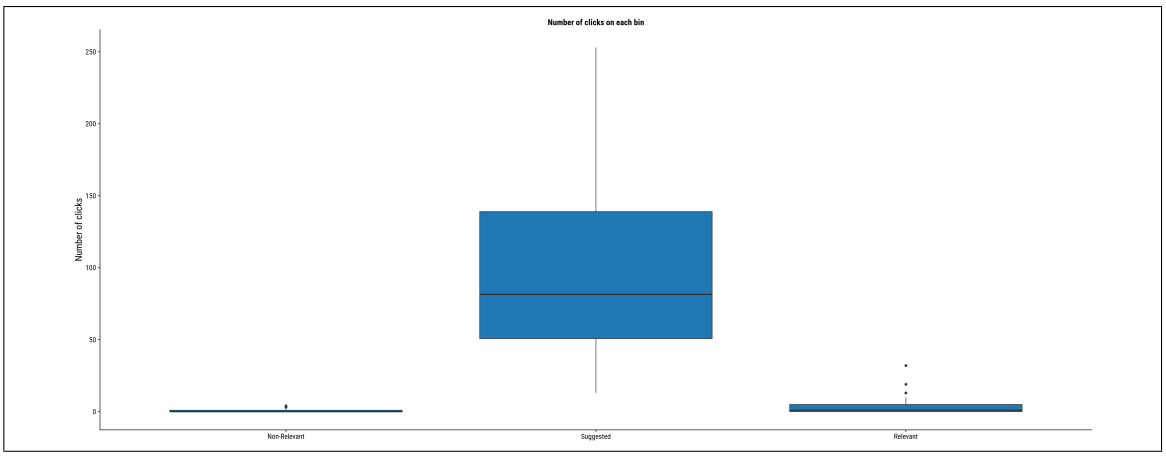

Figure 6.13. Clicks on documents in each bin.

Explore actions were not as common as clicks and lookups. On average a user did 11.93 explore actions ($SE = 2.19$), 158.52 lookup actions ($SE = 14.20$) and 97.14 click actions ($SE = 8.83$). Users did more lookup actions than clicks, which suggests that they stayed for more than 5 seconds scrolling documents' details (see section 5.3).

The interaction with common parts was not different between both interfaces (figure 6.14). Experiments that had the visualization interacted more with the interface. It is important to note that when users had interface without visualization, the space to see the document was bigger. This means that it was not always a scrollable item, so we could not keep track of that action. For this reason, and even though cannot know for sure, it is possible that users read more articles when using the interface without visualization.

Table 6.6. Total actions on the interface

Number of actions	Mean	Std. Dev.	Std. Error
Queries	4.73	3.08	0.46
Total Clicks	97.14	58.60	8.83
Clicks on Suggested Box	92.93	57.59	8.68
Clicks on Relevant Box	3.64	5.98	0.90
Clicks on Non-Relevant Box	0.57	0.97	0.15
Explore	11.93	14.51	2.19
Lookup	158.52	94.19	14.20
Timeline	2.45	3.57	0.54
Regrets	0.80	1.05	0.16
Visualization	25.45	46.71	7.04
Visualization - Explore	13.52	26.08	3.93
Visualization -Clicks	9.70	25.04	3.78

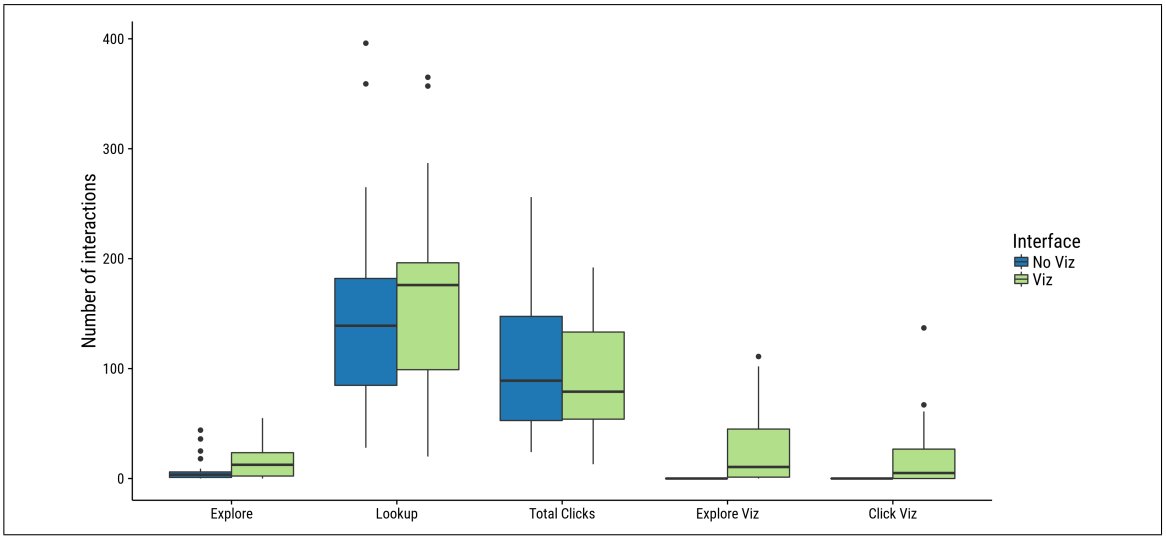

Figure 6.14. Interactions for each interface

6.2.6. Post experiment survey

Results of the post-session study survey are presented in figure 6.15. In terms of satisfaction measured based on expected eventual use of the system (*I would use the system again*, all conditions had an agreement over 75%, with the exception of the condition *Non-viz with BM25*, where only 58% agreed with the statement. These results are correlated

with the answers to the question *I would recommend the system to a colleague*, but the overall agreement is smaller. It is interesting, though, that in terms of *easy of use*, 80% or more people agreed in all conditions, excepting for condition *Viz with Rocchio*, where only 58% agreed with the statement. Comparing these results with those of performance in table 6.4, we can tell that the perception of relevance is closer to the value of precision than to the results of recall at each condition.

They felt the system was easy to use and also said that they would use again. The results for both interfaces were the same, except for the statement *"The system was easy to use"* where the interface without visualization got an average of 4.50 ($SE = 0.16$) and the interface with visualization got 3.68 ($SE = 0.24$); paired t test, $t(21) = 3.4982$, $p = 0.002141$).

Most users felt the system did not miss relevant documents and that it suggested relevant documents. They felt the system was easy to use and also said that they would use again. The results for both interfaces were the same, except for the statement *"The system was easy to use"* where the interface without visualization got an average of 4.5 and the interface with visualization got 3.68.

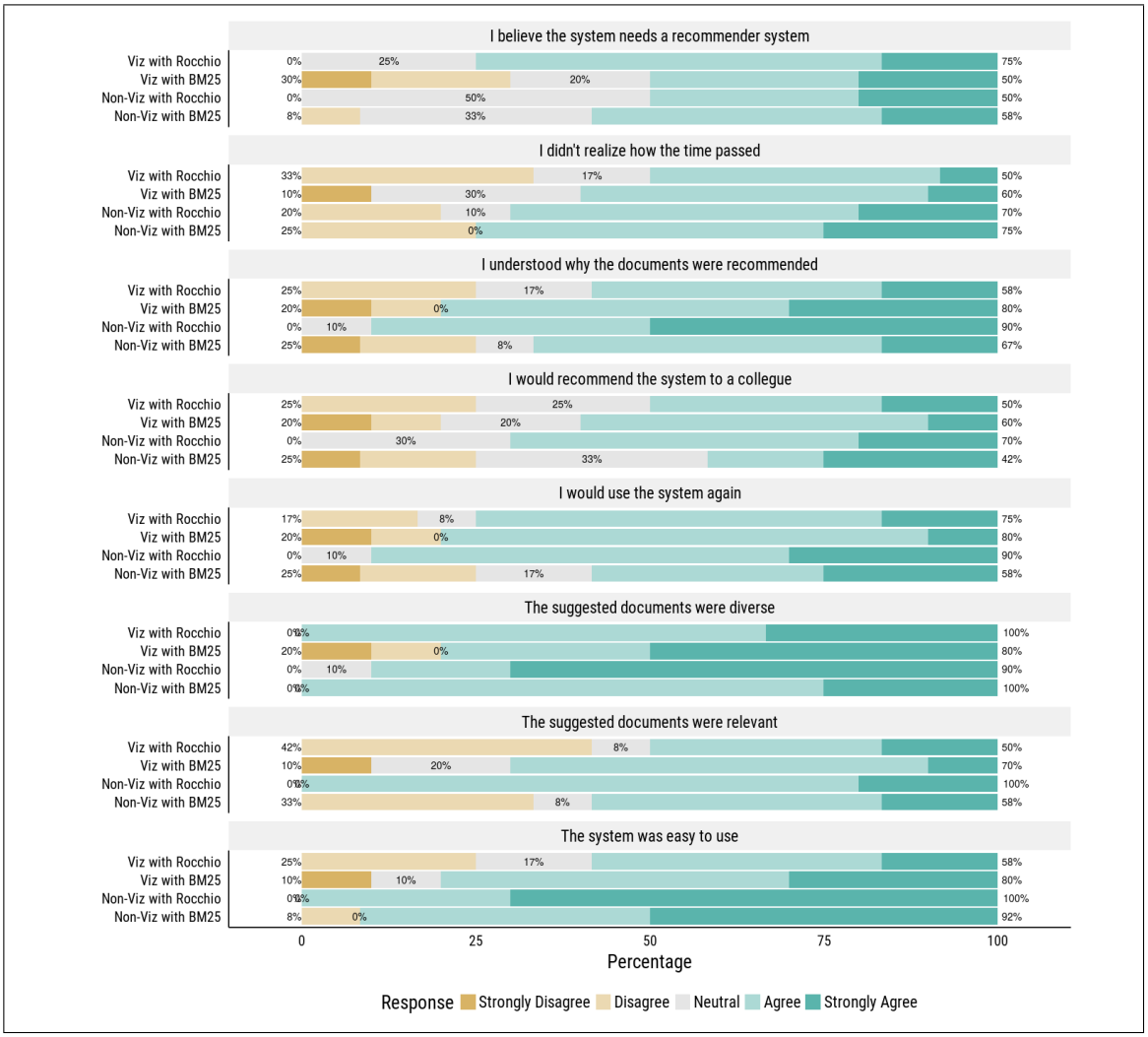

Figure 6.15. Answers to post-session survey, comparing conditions.

6.3. Discussion

In this section we will explore the hypothesis we established in section 3.1.

6.3.1. H1. Does a document visualization affect the task's outcome?

Our results shown in the previous section, specially those in table 6.3 and table 6.4, show a trend towards better recall in visualization conditions when considering the documents seen by users during the session. However, this trend is not the same in terms of precision. To get deeper understanding on this result we analyzed the average accumulated recall by query attempt, shown in figure 6.16. We see that in conditions with visualization, after 6^{th} query attempt, the accumulated recall significantly improves over non-visualization conditions. There is, though, a single outlier user which goes beyond 11 query attempts, and ends up reaching a final recall close to 0.4.

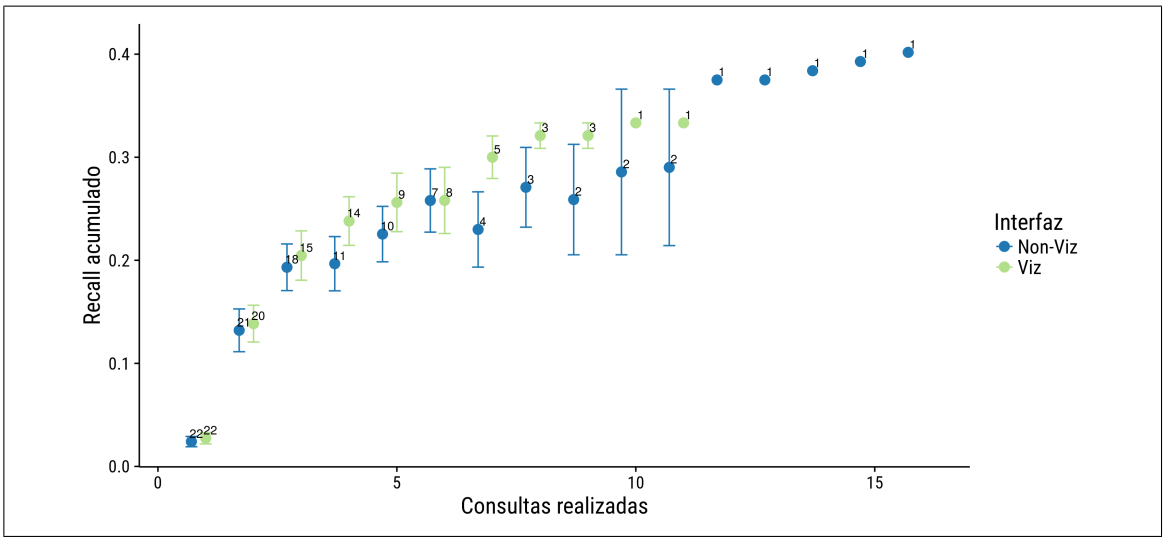

Figure 6.16. Accumulated recall at different queries. Error bar depicts standard error. The number by the circle indicates number of users N.

Comments of users support this result. A user said *"The visualization was useful to find similar studies"* and other said *"I used the visualization to estimate how useful a study could be"*.

6.3.2. H2. Does a relevance feedback algorithm affect the task's outcome?

Our results seem to indicate an interaction effect between algorithm and condition, because Rocchio performs better than BM25 but the difference is significant only with visualization. Performing an analysis considering recall at different query attempts, we show in figure 6.17 that using Rocchio gives better performance at most queries attempts starting from the second attempt. Up to the 8th, the recall of Rocchio is better than recall achieved with BM25, after that, the difference is decreased but the number of users reaching to more than 8 queries is very small (at most 3 users). We again observe an outlier user which, with BM25, performs the best due to persistence, over 15 query attempts.

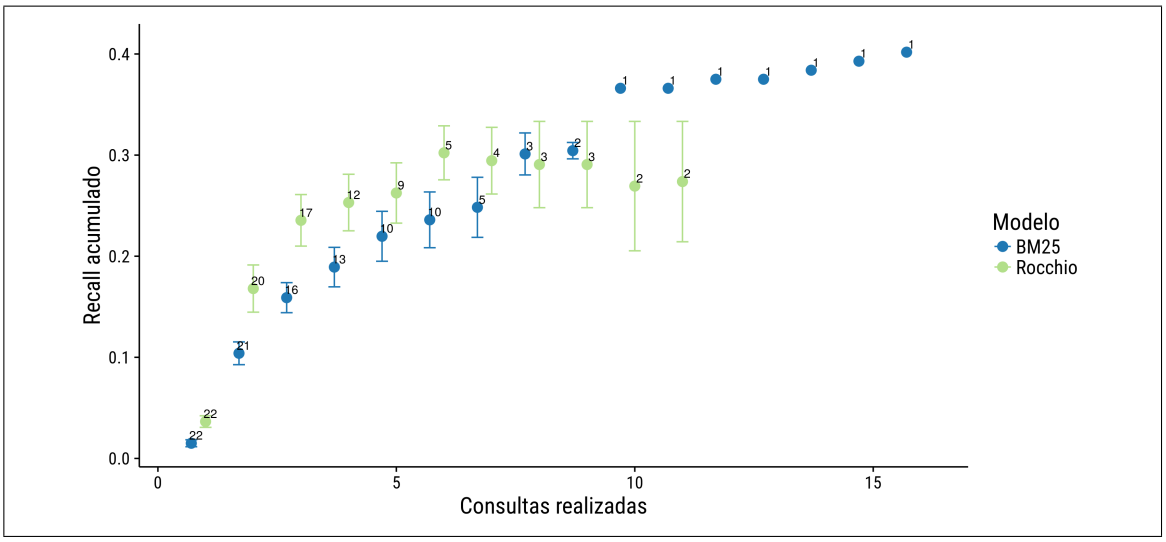

Figure 6.17. Accumulated recall at different queries. Error bar depicts standard error.

6.3.3. H3: Are there other factors that can affect the task's outcome?

To better understand the differences between good and bad results the analysis for this section was perform comparing the users in three groups; *best*, *middle* and *worst*. These groups were computed trying to keep the same number of experiments in each of them. Since the most important metric is recall at session level, we decided to split users' experiments using this value. The limits were:

(i) *worst*: less than 0.12 recall

(ii) *middle*: more than 0.12 and less than 0.30 recall

(iii) *best*: more than 0.30 recall

There are 14 experiments in groups *worst* and *middle* and 16 in *best*.

The experience was measured in three topics: general academic experience, experience working in Evidence Based Health Care. We also measured experience with data visualizations, but we found no differences among groups in this aspect. Having the ability to read research in English does affect recall. Users that strongly agreed to the statement *"I can read research in English"*, had significantly ($p < .01$) better performance than those who did not.

Figure 6.18 shows the experience in Evidence Based Health Care. Having worked in the creation of a Systematic Review helped to get better results ($p < .05$). Users that had created two o more evidence matrices had better recall than those who had never created one or had only created one matrix ($M = 0.23$ vs. $M = 0.13$, $p < .01$).

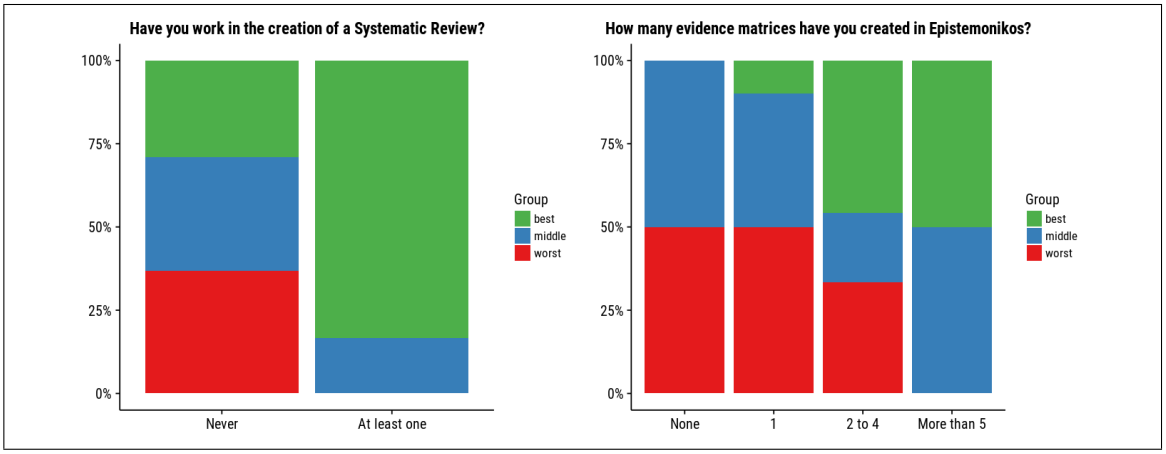

Figure 6.18. Experience with Evidence Based Health Care

7. CONCLUSIONS

In this research we have investigated whether an interactive relevance feedback user interface could help physicians in the process of finding documents to answer a medical question. We have introduced *EpisteAid*, our proposed solution, and we have evaluated it with a user study considering a large dataset and real users of an EBHC system, doctors and medicine students.

We found that the algorithm used in the process is not only relevant for performance metrics, but also for perception of cognitive demand. Rocchio relevance feedback combined with a visualization of documents was found to be better than the other conditions in terms of recall and F-1 score.

We also discovered that reading in English was an important factor. This finding supports the current efforts by Epistemonikos on translating articles into different languages. Experience working in EBHC was also found to be an important variable. This support the need for training physicians to be able to perform in this type of research.

As future work it would be interesting to test other style of algorithms, in particular reinforcement learning, and interfaces that allow users to control weights for keywords and work collaboratively for answering clinical questions.

REFERENCES

Adams, C. E., Polzmacher, S., & Wolff, A. (2013). Systematic reviews: Work that needs to be done and not to be done. *Journal of Evidence-Based Medicine*, *6*(4), 232–235. doi: 10.1111/jebm.12072

Bekhuis, T., & Demner-Fushman, D. (2010). Towards automating the initial screening phase of a systematic review. *Studies in Health Technology and Informatics*, *160*(PART 1), 146–150. doi: 10.3233/978-1-60750-588-4-146

Bekhuis, T., Tseytlin, E., Mitchell, K. J., & Demner-Fushman, D. (2014). Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *PLoS ONE*, *9*(1), 1–10. doi: 10.1371/journal.pone.0086277

Beltran, J. F., Huang, Z., Abouzied, A., & Nandi, A. (2017). Don't just swipe left, tell me why: Enhancing gesture-based feedback with reason bins. In *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 469–480). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/3025171.3025212` doi: 10.1145/3025171.3025212

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python* (1st ed.). O'Reilly Media, Inc.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (1st ed.). springer.

Bostandjiev, S., O'Donovan, J., & Höllerer, T. (2012). Tasteweights: A visual interactive hybrid recommender system. In *Proceedings of the sixth acm conference on recommender systems* (pp. 35–42). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/2365952.2365964` doi: 10.1145/2365952.2365964

Bostock, M., Ogievetsky, V., & Heer, J. (2011, December). D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309. Retrieved from `http://dx.doi.org/10.1109/TVCG.2011.185` doi: 10.1109/TVCG.2011.185

Choi, S., Ryu, B., Yoo, S., & Choi, J. (2012). Combining relevancy and methodological

quality into a single ranking for evidence-based medicine. *Information Sciences*, *214*, 76–90. Retrieved from http://www.sciencedirect.com/science/article/pii/S0020025512003970 doi: http://dx.doi.org/10.1016/j.ins.2012.05.027

Cohen, A. M. (2006). An effective general purpose approach for automated biomedical document classification. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 161–165.

Cohen, A. M., Adams, C. E., Davis, J. M., Yu, C., Yu, P. S., Meng, W., ... Smalheiser, N. R. (2010). Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. *Proceedings of the ACM international conference on Health informatics - IHI '10*, 376. Retrieved from http://portal.acm.org/citation.cfm?doid=1882992.1883046 doi: 10.1145/1882992.1883046

Cormen, T. H. (2009). *Introduction to algorithms*. MIT press.

di Sciascio, C., Sabol, V., & Veas, E. E. (2016). Rank as you go: User-driven exploration of search results. In *Proceedings of the 21st international conference on intelligent user interfaces* (pp. 118–129). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2856767.2856797 doi: 10.1145/2856767.2856797

Elliott, J. H., Turner, T., Clavisi, O., Thomas, J., Higgins, J. P. T., Mavergames, C., & Gruen, R. L. (2014). Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. *PLoS Medicine*, *11*(2), e1001603. Retrieved from http://dx.plos.org/10.1371/journal.pmed.1001603 doi: 10.1371/journal.pmed.1001603

Elsevier. (2016). Embase, biomedical evidence is essential. [Computer software manual]. Retrieved 2016-10-14, from http://store.elsevier.com/embase

Engel, D., Hüttenberger, L., & Hamann, B. (2012). A survey of dimension reduction methods for high-dimensional data analysis and visualization. In *Oasics-openaccess series in informatics* (Vol. 27).

Epistemonikos. (2016). Epistemonikos database methods [Computer software manual]. Retrieved 2016-10-14, from http://www.epistemonikos.org/en/about_us/methods

Felizardo, K. R., Andery, G. F., Paulovich, F. V., Minghim, R., & Maldonado, J. C. (2012). A visual analysis approach to validate the selection review of primary studies in systematic reviews. *Information and Software Technology*, *54*(10), 1079–1091. Retrieved from http://dx.doi.org/10.1016/j.infsof.2012.04.003 doi: 10.1016/j.infsof.2012.04.003

Felizardo, K. R., Souza, S. R. S., & Maldonado, J. C. (2013). The use of visual text mining to support the study selection activity in systematic literature reviews: A replication study. In *Proceedings - 2013 3rd international workshop on replication in empirical software engineering research, reser 2013.* doi: 10.1109/RESER.2013.9

García Adeva, J. J., Pikatza Atxa, J. M., Ubeda Carrillo, M., & Ansuategi Zengotitabengoa, E. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, *41*(4 PART 1), 1498–1508. doi: 10.1016/j.eswa.2013.08.047

Higgins JPT, G. S. (2011). *Cochrane handbook for systematic reviews of intereventions, version 5.1.0 [updated march 2011]* (5th ed.). The Cochrane Collaboration. Retrieved from http://www.cochrane-handbook.org

Hijikata, Y., Kai, Y., & Nishida, S. (2012). The Relation between User Intervention and User Satisfaction for Information Recommendation. In (pp. 2002–2007).

Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., ... Thayer, K. (2016). SWIFT-Review: a text-mining workbench for systematic review. *Systematic reviews*, *5*, 87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/27216467{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4877757 doi: 10.1186/s13643-016-0263-z

Jones, K. S., Walker, S., & Robertson, S. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing & Management*, *36*(6), 779 - 808. Retrieved from http://www.sciencedirect.com/science/article/pii/S0306457300000157 doi: http://dx.doi.org/10.1016/S0306-4573(00)00015-7

Jonnalagadda, S., & Petitti, D. (2013). A new iterative method to reduce workload in systematic review process. *International journal of computational biology and drug design*, *6*(1-2), 5–17. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84874422148{&}partnerID=tZOtx3y1 doi: 10.1504/IJCBDD.2013.052198

Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the Royal Society of Medicine*, *96*(3), 118–121. doi: 10.1258/jrsm.96.3.118

Kilicoglu, H., Demner-Fushman, D., Rindflesch, T. C., Wilczynski, N. L., & Haynes, R. B. (2009). Towards Automatic Recognition of Scientifically Rigorous Clinical Research Evidence. *Journal of the American Medical Informatics Association*, *16*(1), 25–31. doi: 10.1197/jamia.M2996

Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., & Kobsa, A. (2012). Inspectability and control in social recommenders. In *Proceedings of the sixth acm conference on recommender systems* (pp. 43–50). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2365952.2365966 doi: 10.1145/2365952.2365966

Knijnenburg, B. P., & Willemsen, M. C. (2011). Each to His Own : How Different Users Call for Different Interaction Methods in Recommender Systems. *Proceedings of the 5th ACM conference on Recommender systems - RecSys '11*, 141–148. Retrieved from http://dl.acm.org/citation.cfm?id=2043932.2043960 doi: 10.1145/2043932.2043960

Kouznetsov, A., & Japkowicz, N. (2010). Using Classifier Performance Visualization to Improve Collective Ranking Techniques for Biomedical Abstracts Classification. In A. Farzindar & V. Kešelj (Eds.), *Advances in artificial intelligence: 23rd canadian conference on artificial intelligence, canadian ai 2010, ottawa, canada, may 31 – june 2, 2010. proceedings* (pp. 299–303). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-13059-5{_}33 doi: 10.1007/978-3-642-13059-5_33

Kouznetsov, A., Matwin, S., Inkpen, D., Razavi, A. H., Frunza, O., Sehatkar, M., ...

O'Blenis, P. (2009). Classifying Biomedical Abstracts Using Committees of Classifiers and Collective Ranking Techniques. In Y. Gao & N. Japkowicz (Eds.), *Advances in artificial intelligence: 22nd canadian conference on artificial intelligence, canadian ai 2009 kelowna, canada, may 25-27, 2009 proceedings* (pp. 224–228). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-01818-3{_}29 doi: 10.1007/978-3-642-01818-3_29

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York, NY, USA: Cambridge University Press.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th python in science conference* (pp. 51–56).

Miwa, M., Thomas, J., O'Mara-Eves, A., & Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*, *51*, 242–253. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1532046414001439 doi: 10.1016/j.jbi.2014.06.005

O Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, *4*(1), 5. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4320539{&}tool=pmcentrez{&}rendertype=abstract doi: 10.1186/2046-4053-4-5

Olorisade, B. K., de Quincey, E., Brereton, P., & Andras, P. (2016). A critical analysis of studies that address the use of text mining for citation screening in systematic reviews. In *Proceedings of the 20th international conference on evaluation and assessment in software engineering* (pp. 14:1–14:11). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2915970.2915982 doi: 10.1145/2915970.2915982

Otto, M., & Thornton, J. (2016). Bootstrap [Computer software manual]. Retrieved

2016-08-01, from http://getbootstrap.com/

Parra, D., & Brusilovsky, P. (2015, June). User-controllable personalization. *Int. J. Hum.-Comput. Stud.*, *78*(C), 43–67. Retrieved from http://dx.doi.org/10.1016/j.ijhcs.2015.01.007 doi: 10.1016/j.ijhcs.2015.01.007

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Peltonen, J., Strahl, J., & Floréen, P. (2017). Negative relevance feedback for exploratory search with visual interactive intent modeling. In *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 149–159). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/3025171.3025222 doi: 10.1145/3025171.3025222

Rada, G., Perez, D., & Capurro, D. (2013). Epistemonikos: a free, relational, collaborative, multilingual database of health evidence. *Studies in health technology and informatics*, *192*, 486–490. doi: 10.3233/978-1-61499-289-9-486

Sackett, D. L., & Rosenberg, W. M. (1995). The need for evidence-based medicine. *Journal of the Royal Society of Medicine*, *88*(11), 620–624.

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *Bmj*, *312*(7023), 71–72.

Salton, G. (1971). *The smart retrieval system—experiments in automatic document processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for informatio nvisualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, 336–343. doi: 10.1109/VL.1996.545307

Thomas, J., McNaught, J., & Ananiadou, S. (2011). Applications of text mining within systematic reviews. *Research Synthesis Methods*, *2*(1), 1–14. Retrieved from http://doi.wiley.com/10.1002/jrsm.27 doi: 10.1002/jrsm.27

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011, March). The numpy array: A structure for efficient numerical computation. *Computing in Science Engineering*, *13*(2),

22-30. doi: 10.1109/MCSE.2011.37

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. a. (2010). Modeling Annotation Time to Reduce Workload in Comparative Effectiveness Reviews Categories and Subject Descriptors Active Learning to Mitigate Workload. *Proceedings of the 1st ACM International Health Informatics Symposium. ACM,*, 28–35. doi: 10.1145/1882992.1882999

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. In *Proceedings of the 2nd acm sighit international health informatics symposium* (pp. 819–824). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2110363.2110464 doi: 10.1145/2110363.2110464

Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. a. (2010). Active Learning for Biomedical Citation Screening. *Kdd2010*, 173–181. doi: 10.1145/1835804.1835829

Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2011). Who Should Label What? Instance Allocation in Multiple Expert Active Learning. *Sdm*, 176–187. Retrieved from http://epubs.siam.org/doi/abs/10.1137/1.9781611972818.16 doi: 10.1137/1.9781611972818.16

Wallace, B. C., Trikalinos, T. a., Lau, J., Brodley, C., & Schmid, C. H. (2010). Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, *11*(55), 55. doi: 10.1186/1471-2105-11-55

Yu, W., Clyne, M., Dolan, S. M., Yesupriya, A., Wulf, A., Liu, T., ... Gwinn, M. (2008). GAPscreener: An automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. *BMC Bioinformatics*, *9*, 205. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2387176/ doi: 10.1186/1471-2105-9-205

APPENDIX

A. PRE STUDY SURVEY

The survey was given in spanish. The following sections

1. Please read each statement and indicate to what extend you agree or disagree (strongly disagree, disagree, neutral, agree, strongly agree) with each of them.

Table A.1. Question 1 pre-study survey

	Statement
1	I know what an automatic classifier is how it works
2	I usually work with data visualizations
3	I quickly understand data visualizations
4	I can read research articles of my area in English without problems

2. Please read each statement and indicate to what extend you agree or disagree (strongly disagree, disagree, neutral, agree, strongly agree) with each of them.

Table A.2. Question 2 pre-study survey

	Statement
1	I know therapeutic indications for venous thromboembolism during pregnancy
2	I know the risks of using low molecular weight heparins during pregnancy
3	I have extensive knowledge about anticoagulation during pregnancy
4	I feel capable to recommend or advise a colleague about the real benefits and risks and costs of using Echinacea for colds
5	I know the interventions that are helpful during a cold

3. How many evidence matrices have you created?
   - None
   - Just one

- Two to four
- More than five

4. Have you participated in the creation of a systematic review?
    - Non
    - yes
5. You are a ...?
    - Medicine Student
    - Physician
6. If you have finished your studies, please indicate your specialization

B. POST STUDY SURVEY

The survey was given in spanish. The following sections

1. Please read each statement and indicate to what extend you agree or disagree (strongly disagree, disagree, neutral, agree, strongly agree) with each of them.

Table B.1. Question 1 post-study survey

	Statement
1	I understood why the documents were recommended for the topic
2	The documents suggested by the system seemed relevant to the topic of the question
3	The documents suggested by the system were diverse
4	I quickly felt familiar with the interface
5	I felt the system was easy to use
6	I didn't realize how time passed while using the recommendation interface
7	The system made feel that it didn't miss relevant documents for the search
8	I would use the system again to search for documents
9	I would recommend the system to a colleague
10	I think the system requires an automatic recommendation system

2. Only for experiments that used the visualization. Please read each statement and indicate to what extend you agree or disagree (strongly disagree, disagree, neutral, agree, strongly agree) with each of them.

Table B.2. Question 2 post-study survey. Only for experiments with visualization

	Statement
1	The visualization increased my satisfaction with the suggested documents
2	The visualization was useful to pick what documents to classify next
3	I have extensive knowledge about anticoagulation during pregnancy
4	The visualization help to understand why the documents were recommended
5	The visualization colors were appropriate to understand

3. In general, what was the process you followed to decide whether a document was relevant or not? What parts of the interface you used to make a decision?
4. Do you have any comments on the activity?

C. CONSENT FORM

DECLARACIÓN DE INFORMACIÓN PARA EL PARTICIPANTE

Proyecto de Investigación:
Estudiando Factores Humanos para Entender la Aceptación de Recomendaciones por parte de Usuarios.
(Studying Human Factors to Understand User Acceptance in Recommender Systems)

(1) ¿De qué se trata el estudio?

El estudio consiste en investigar la efectividad de diferentes técnicas de cálculo y visualización de recomendaciones automatizadas y personalizadas.

(2) ¿Quién está desarrollando el estudio?

Director de la investigación: Denis Alejandro Parra Santander, Departamento de Ciencia de la Computación, Escuela de Ingeniería, Pontificia Universidad Católica de Chile. Vicuña Mackenna 4860, Macul, Santiago. Fono: (562) 354-4442. Correo electrónico: dparra@ing.puc.cl

(3) ¿Qué involucra el estudio?

El estudio involucra sentarse frente a un computador, responder un cuestionario con preguntas demográficas y de nivel de expertis al empezar el estudio, interactuar con una interfaz gráfica que recomienda items de forma personalizada, y luego responder un cuestionario al final del estudio.

(4) ¿Cuánto tiempo durará el estudio?

Entre 45 minutos y 1 hora y media

(5) ¿Puedo retirarme anticipadamente del ejercicio?

Sí

(6) ¿Quiénes conocerán los resultados?

El investigador responsable, los alumnos de pre y postgrado que trabajen en el análisis de datos. De forma agregada (métricas descriptivas estadísticas y de desviación de datos), quienes lean las publicaciones resultantes del estudio. Es importante recalcar que a nivel de publicaciones los datos serán anonimizados y se presentarán de forma agregada, de manera que no sea posible identificar a una persona solamente por las respuestas que entregó durante el estudio.

(7) ¿Quién es responsable de todos los registros y datos?

El investigador principal.

(8) ¿Me beneficiará en algo la participación en este estudio?

Podría ayudarte a descubrir items que no conoces en el área en la que se realicen las recomendaciones, ya sea musicales, de libros, películas o artículos científicos.

(9) ¿Puedo contarle a otras personas sobre el estudio?

Si.

(10)¿Qué debo hacer si necesito más información?

Cuando haya leído esta información, la persona a cargo de administrar la entrevista responderá cualquier pregunta que usted pueda tener. Si desea profundizar en algún aspecto del estudio, Ud. puede contactar directamente a Denis Alejandro Parra Santander al (02) 354-4442 o por correo electrónico a dparra@ing.puc.cl

(11) ¿Qué pasa si tengo alguna queja o inquietud?

Cualquier persona con inquietudes o quejas sobre la conducta de un estudio de investigación puede ponerse en contacto con el Comité de Ética de la Escuela de Ingeniería de la P. Universidad Católica de Chile, representado por el Sr. Juan Enrique Coeymans, Presidente del Comité de Ética, Av. Vicuña Mackenna 4860, Santiago, al teléfono (02) 354-1189 o por correo electrónico a la dirección jec@ing.puc.cl

No firme la presente carta hasta que haya leído toda la información proporcionada y haya hecho todas las preguntas que desee. Se le proporcionará copia de este documento.

FORMULARIO DE CONSENTIMIENTO INFORMADO

Yo, ... , doy mi consentimiento para participar en el estudio sobre "Técnicas de visualización en interfaces de recomendación" que forma parte del proyecto de investigación "Estudiando Factores Humanos para Entender la Aceptación de Recomendaciones por parte de Usuarios", llevada a cabo por la Facultad de Ingeniería de la Pontificia Universidad Católica de Chile

Al dar mi consentimiento, yo reconozco que:

1. Se me han explicado todos los procedimientos y el tiempo requerido para participar en la encuesta, y toda pregunta sobre el proyecto ha sido respondida a mi entera satisfacción.

2. He leído la Declaración de Información para el Participante y se me ha ofrecido la oportunidad de examinar toda la información sobre mi participación en el proyecto.

3. Entiendo que puedo retirarme de la entrevista en cualquier momento, sin que ello afecte mi relación con el investigador(a) ahora o en el futuro.

4. Entiendo que mi participación es estrictamente confidencial y que ninguna información que revele mi identidad será utilizada en modo alguno.

5. Entiendo que mi participación en esta entrevista es completamente voluntaria – no estoy bajo ninguna presión para participar ni entregar mi consentimiento.

6. Se podrán registrar las entrevistas en audio y/o video.

7. Entiendo que si no quisiera continuar puedo retirarme de la entrevista en cualquier momento. Cualquier información que pude haber dado al entrevistador hasta ese momento será destruida.

Firma: ……

Nombre:

Fecha:

Investigador Responsable	Presidente Comité de Ética Escuela de Ingeniería
Denis Parra Santander, tel (02) 354-4442 correo electrónico: dparra@ing.puc.cl	Sr. Juan Enrique Coeymans, tel (02) 354-1189 correo electrónico: jec@ing.puc.cl

Este documento se firma en duplicado quedando una copia para el participante