Practical Considerations when Filtering Documents

Desmond Elliott*
School of Informatics
University of Edinburgh
d.elliott@ed.ac.uk

Leif Azzopardi
School of Computing Science
University of Glasgow
leif@dcs.gla.ac.uk

ABSTRACT

Implementing, configuring, and running an information filtering system in a practical setting is a difficult and challenging problem. This is due to variety and configuration of available system components along with additional factors such as topic length, feedback, and system training. Moreover, the interplay between the different components and additional factors can lead to degraded system performance when adding or manipulating particular components. We explore the interactions and effects of different components and some of the factors with respect to performance. The main contribution of this paper is a better understanding of how to configure filtering systems along with the possible pitfalls of applying conflicting components which harm performance and result in a poor user experience.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval

General Terms: Performance, Experimentation

Keywords: Filtering, Selective Dissemination

1. INTRODUCTION

The growth of information available on the internet can lead to the problem of *information overload*, which can be defined as "represent[ing] a state of affairs where an individual's efficiency in using information in their work is hampered by the amount of relevant, and potentially useful, information available to them."[5]. One approach to alleviating this problem for long-term information needs is to delegate part of the information seeking process to an *information filtering system*. Settings where a filtering system could be useful include: matching news stories with personal interests; finding journal articles and conference papers for academics; and helping organisations find consumer and press opinions on new products. A filtering system typically matches documents from an incoming document stream against a set of topic profiles with the aim of maximising the accuracy of the documents delivered to each topic profile. Designing and implementing an effective filtering system is a difficult task because of the dynamic nature of incoming content and the evolving nature of the user's long term information needs [11]. While these external factors contribute to the difficulty of the problem, there are a number of potentially conflicting system-based factors that further complicate the problem. These include: (i) the representation of topics and documents [3, 6], (ii) the scoring function used to determine the relationship between a document and a topic [7, 8, 9, 12]; (iii) the threshold adaptation method used to select documents [4]; and (iv) the topic adaptation method used to incorporate implicit and explicit feedback [1].

These factors have generally been examined in isolation of each other, so it is not possible to categorically state the impact of the interplay of these factors on system performance. For example, it has been found that increasing the filtering threshold tends to improve precision at the expense of recall, but increasing the topic length tends to improve recall at the expense of precision. *But, what happens when we do both?* This paper attempts to explore what happens when these methods are used in combination and undertakes an empirical study that investigates the influence of each factor and the interplay between these factors on filtering performance. The main contribution of this paper is a working guide for both researchers and practitioners about the cause and effect of the different components within a filtering system, along with an improved understanding of the filtering process.

2. INFORMATION FILTERING

A typical information filtering system consists of a matching function, a threshold adaptation method, and a topic

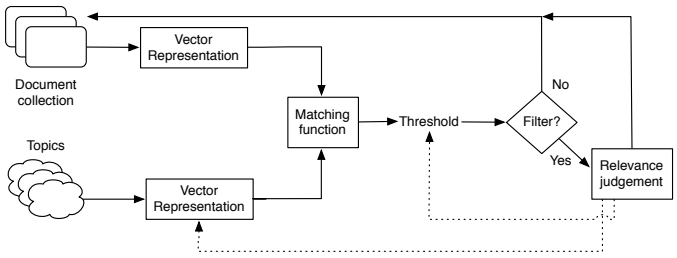


Figure 1: A schematic of a typical information filtering system. The dotted lines represent activity associated with threshold and/or topic adaptation.

*Work carried out while the author was at the University of Glasgow.

IIiX 2012, Nijmegen, The Netherlands

adaptation method (see the TREC Filtering Track reports for numerous examples of such systems [7, 8, 9, 12]). Figure 1 provides a schematic of such an information filtering system. The incoming documents are usually represented as a document vector after tokenisation, stemming, and stop word removal. Each document representation is then compared against a set of topic profiles via the matching function. Documents exceeding the filtering threshold are considered to be relevant to the topic, and are filtered and presented to the user for judgement, otherwise they are discarded. The result of the filtering decision and the subsequent judgement can be exploited to update the topic profile itself or score distributions for relevant and irrelevant documents. System performance can be measured in terms of *precision* and *recall*, where precision is the ratio of correctly filtered documents to incorrectly filtered documents and recall is the ratio of correctly filtered documents to all possible correct documents, on a topic-by-topic basis.

Matching Functions: The dominant approaches are vector-space variants and probabilistic variants. The most popular and effective model is the probabilistic model derived from the Okapi BM25 scoring function, which has been extensively evaluated at TREC [7, 8, 9, 12]. Documents and topics are represented as weighted term-vectors in a similar manner as within retrieval systems, and weighting functions such as TF-IDF to match documents with topic profiles are employed.

Threshold Adaptation Methods: This component attempts to alter the filtering threshold for each topic based on characteristics of the previously filtered documents [1]. The motivation behind threshold adaptation is that some topics have poor precision with a static threshold, as a result of filtering too many irrelevant documents. In this case, threshold adaptation increases the filtering threshold with the aim of increasing topic precision. In [2], filtering thresholds for each topic were set to halfway between the scores of the relevant and irrelevant documents in the training set. This method increases the threshold over the course of filtering, and also the precision. Other state of the art methods use score distributions to adapt the threshold [4, 13].

Topic Adaptation Methods: The goal of topic adaptation is to learn which terms are most representative of a topic as filtering progresses. This can take the form of query expansion, given the result of relevance judgements on presented documents, or updating the weights of existing terms in a topic profile, or both. Topic adaptation aims to increase recall by building a better representation of the long-term interests [1].

Conflicting Interests between Components

Most information filtering systems use both threshold and topic adaptation methods, however, the interplay between these methods provides a potentially conflicting arrangement. Threshold adaptation increases precision at the expense of recall, while topic adaptation increases recall at the expense of precision. One possible outcome of using these methods at the same time is that both precision and recall are decreased. Furthermore, the matching functions used in filtering systems are often directly ported from retrieval systems, so another confounding factor may be that the function is unbounded (i.e may produce a score from zero to infinity). When using an unbounded scoring function, any increase in topic length will mean a higher document scores are obtained for subsequent documents. This is because a longer topic profile will have more terms to match with the document, which will have the effect of increasing recall at the expense of precision because many more documents will exceed the filtering threshold. In this paper, we also consider a bounded scoring function to isolate the effect of increasing scores and compare it against the unbounded scoring function. Also related to topic length and scoring functions is the issue of topic adaptation via query expansion, which can increase the number of terms used to represent the topic. As a result the scores assigned to subsequent documents are likely to be higher if the score of a document depends on the number of query terms. This will implicitly raise recall because more documents will exceed the threshold. The combination of these factors is likely to lead to a substantial decrease in precision if threshold adaptation is not employed to counter-balance this effect.

3. METHODOLOGY

We examined the influence of: (i) different filtering thresholds, (ii) different topic lengths, (iii) a bounded or unbounded matching function, (iv) threshold adaptation using the midpoint method, and (v) topic adaptation using relevance feedback and the Rocchio method. The experimental filtering system followed the structure of Figure 1 and was written in C++ using the Lemur Toolkit[1].

The data used in this study compromised of the following TREC collections and topics from the Filtering Tracks: FBIS, AP, FT, and RCV-1. These ranged in size from 130k to 806k documents. We use the TREC WSJ collection as a reference collection to provide estimates of term statistics, such as IDF. We reserved the first 10% of each collection for training data. The remaining 90% of each collection was then used for testing. Documents were indexed in document identifier order, and stemmed using the Porter Stemmer and stop words are removed from the documents using the stop word list accompanying the toolkit.

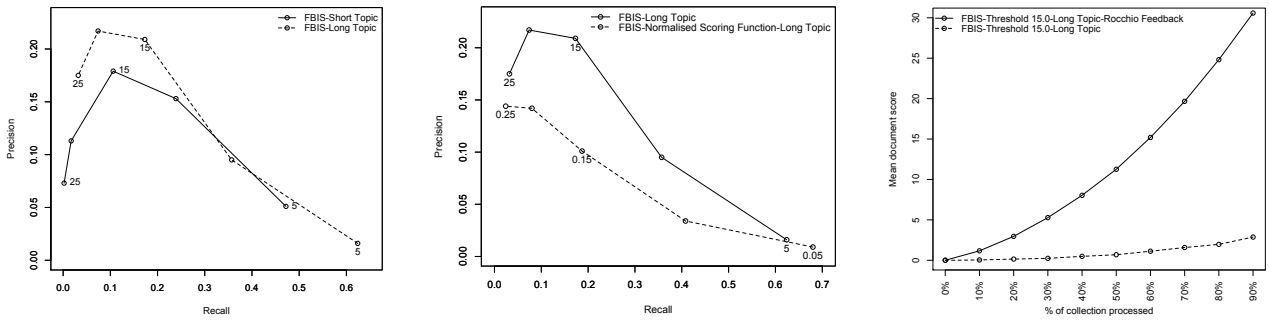
We use the vector-space model to represent documents and topics. We chose this because it was the most commonly adopted model in previous filtering systems and many of the state-of-the-art methods are based on this model and used this in conjunction with the Okapi BM 25 scoring function [10]. The parameters of the scoring function were set to $k_1 = 2.0$ and $b = 0.75$, and $k_3 = 1.2$. We also use a normalised version of Okapi BM25 to create a bounded version of the scoring function.

Topics were represented as a weighted-term vector in either a *short* or *long* length. The short topic profiles were constructed from the TREC Title, with a mean length 2.7 terms, and from the TREC Title and Description fields for long topics, with a mean length 5.3 terms. The weight of each term in the topic representation is calculated using *tf-idf* where the Inverse Document Frequency (*idf*) value was calculated using the document frequency data using WSJ[2].

Threshold adaptation was performed using the mid-point of the relevant document scores and irrelevant document scores identified in the training set [2]. When used, the topic representations were adapted through positive feedback using the Rocchio method with parameters set to $\alpha = 1.0$, $\beta = 0.75$, & $\gamma = 0$.

[1] http://www.lemurproject.org

[2] If a term does not exist in the WSJ collection, the idf component is assigned a value of 1 because it appeared in at least the topic.



(a) Topic length and threshold. (b) Bounded or unbounded scoring function. (c) Score evolution.

Figure 2: Precision and recall as thresholds vary for different factors (left, mid.), Score Evolution (right).

4. EMPIRICAL FINDINGS

We present the effect of each factor on precision and recall of filtered documents and also present the interplay between topic and threshold adaptation. We are unable to report all results due to space constraints[3].

Varying filtering threshold: We fix the scoring function to *unbounded Okapi BM25*, and apply no topic or threshold adaptation during the filtering process. The effect of varying the initial filtering threshold from 5.0 to 25.0 can be seen in Figure 2(a) for short and long topic profiles. As the threshold increases, recall decreases, tending to zero. However, the relationship with precision is more complicated: as the threshold increases, this initially leads to an increase in precision, but after a certain point the threshold becomes far to strict and precision also tends to zero. This trend was found to occur in all document collections and was regardless of the choice of a bounded or unbounded scoring function.

Short or Long topic length: In Figure 2(a), we can also see the difference between the short topic profiles (solid line) and longer profiles (dashed line). At the lowest threshold, we see that the long topic profiles have greater recall, but lower precision. At each threshold point, recall is higher when topic length is long. However, as the threshold is increased, the long topics also yield greater precision (see threshold of 15, for example.). This is because these longer topics are manually crafted by the user and provide more matching terms with which to identify relevant documents. When terms are selected automatically, via topic adaptation, this also increases topic length, but performance did not improve. However, for precision, we observe that better descriptions of the initial information need tends to result in greater levels of precision as the threshold increases.

Scoring function: Figure 2(b) shows the performance of the bounded and unbounded function when topic length is set to long. We can see that using a normalized scoring function does not substantially affect precision at high thresholds. However, if we continue to increase the threshold, precision and recall would tend to zero. Overall, the unbounded scoring function provides superior precision and recall over the bounded scoring function across each level of recall. This is perhaps due to the effects the bounding has upon the score distributions, but further work is required to determine the exact cause.

Threshold adaptation: To examine the influence of threshold adaptation we used the *midpoint method* and started with the initial thresholds of 5 to 25. In Figure 3, we can see the results when topic length was long, when there was threshold adaptation (dashed line) and without threshold adaptation (solid line). At lower thresholds, precision is dramatically improved (2% to 14.9%), at the expense of recall (63% down to 19%). However, at higher initial thresholds, the precision is generally improved, with smaller losses to recall, for example when the threshold starts at 25, precision increased by 3%, while recall is almost unchanged. To develop an understanding of why recall decreases like this, we tracked the mean threshold throughout the filtering process and note that the threshold under the midpoint method always increases (an intuition put forward by Allan et al. [1]), and the increase in threshold is less pronounced at higher thresholds, than at lower thresholds. A potential solution to mitigating this trend is to use the scores of all documents instead of all filtered documents or by decaying the contribution of older document scores. Overall, though, we observed that threshold adaptation enhances precision by trading off recall.

Topic adaptation: The effect of topic adaptation is also shown in Figure 3 for long topic representations (dotted line). The precision of the filtered documents is almost zero, while recall remained relatively high across thresholds. To provide an explanation why this is the case, Figure 2(c) shows the mean document score during the course of filtering. We can see that there is an almost exponentially increase to the mean document score when topic adaptation is used on the unbounded scoring function. After processing 60% of the document collection, a threshold of 15 would recommend nearly all subsequent documents (which accounts for the loss in precision). It would appear that topic adaptation requires moderation by threshold adaptation to work effectively.

Topic & Threshold Adaptation: Finally, Figure 3 shows the interplay of simultaneous topic and threshold adaptation on performance (dot-dash line). While the precision is improved over topic adaptation, this is at the expense of recall. However, when compared to threshold adaptation alone, the precision is substantially lower for similar recall. What we can observe is that the intention of topic adaptation to increase recall and threshold adaptation to increase precision can have negative consequences. The components may not work together, and thus be harmful employing a fixed threshold or threshold adaptation.

[3]A full technical report is available [14].

5. DISCUSSION & FUTURE WORK

Factor/Component and Observations
Varying Threshold As the threshold ↑, precision ↑ and recall ↓. This is observed until a point of maximum precision, after which both precision ↓ and recall ↓. With larger increases using the bounded function.
Topic Length As the initial topic length ↑, precision ↑ and recall ↑. With larger increases using the bounded function.
Midpoint Threshold Adaptation As the threshold ↑, precision ↑ and recall ↓.
Rocchio Topic Adaptation With further topic adaption (and subsequent increase to topic length) ↑, precision ↓ and recall ↑.
Threshold & Topic Adaptation As threshold ↑ and topic length ↑, precision ↓ and recall ↓.

Table 1: Summary of findings of the effect of components of precision and recall on a filtering system.

We have studied the influence of different components on precision and recall during the filtering process. We have shown that there are major difficulties in configuring such systems such that the performance is optimized. In particular, we have shown that using the midpoint threshold adaptation method alongside topic adaptation is detrimental to the overall system performance - and thus care must be taken when selecting and configuring components. From our study, we have tried to summarize our major findings and present them in Table 1 to serve as a helpful guide to practitioners and researchers of filtering systems.

However, we note that this study has a number of limitations. We have only employed a subset of the state of the art methods for topic and threshold adaptation. It may be that other methods work better together, but this is left for future investigation. Also, in other work, it has been shown that topic adaptation can improve performance. We highlight that we have only used 10% of the data for training, whereas other work used significantly more training data [4, 13]. In fact, using a large proportion of the data for supervised machine learning tasks is commonplace and would allow for a test set within the training set for parameter tuning. Another point is the term selection strategy for choosing expansion terms. While the longer initial topics performed substantially better than the short topics, automatically expanded topics performed poorly. Topic length and term selection are two potentially confounding factors that need to further examined. These limitations are quite general, and do not just pertain to our work on filtering, but highlight a number of major technical challenges for developing robust, reliable and usable filtering systems. These are: (i) developing methods that utilize only a small amount of training data, (ii) improving the automatic term selection algorithms and process, (iii) understanding the interplay between the scoring algorithm and topic length, (iv) developing thresholding methods that account for this interplay.

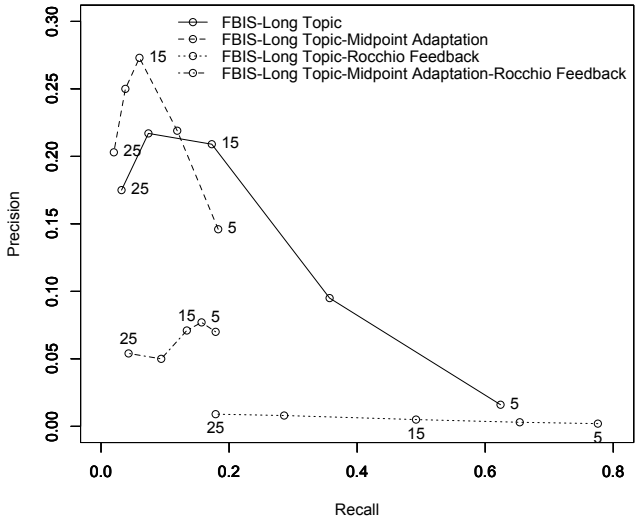


Figure 3: The effect of threshold and topic adaptation on Precision and Recall. The labels show the initial threshold values.

6. REFERENCES

[1] J. Allan. Incremental relevance feedback for information filtering. In *SIGIR '96*, pages 270–278, 1996.

[2] J. Allan, J. P. Callan, W. B. Croft, L. Ballesteros, D. Byrd, R. C. Swan, and J. Xu. INQUERY does battle with TREC-6. In *Text REtrieval Conference*, pages 169–206, 1997.

[3] G. Amati, D. D'Aloisi, V. Giannini, and F. Ubaldini. A framework for filtering news and managing distributed data. *Journal of Uni. Comp. Sci.*, 3:1007–1021, 1997.

[4] A. T. Arampatzis and A. van Hameren. The score-distributional threshold optimization for adaptive binary classification tasks. In *SIGIR '01*, pages 285–293, 2001.

[5] D. Bawden and L. Robinson. The dark side of information. *J. of Info. Sci.*, 35(2):180–191, 2009.

[6] J. Callan. Learning while filtering documents. In *SIGIR '98*, pages 224–231, 1998.

[7] D. A. Hull. The TREC-6 Filtering Track. In *The Sixth Text REtrieval Conference*, 1997.

[8] D. A. Hull. The TREC-7 Filtering Track. In *The Seventh Text REtrieval Conference*, pages 33–56, 1999.

[9] D. A. Hull and S. Robertson. The TREC-8 Filtering Track. In *The Eighth Text REtrieval Conference*, 2000.

[10] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval. *Inf. Process. Manage.*, 36(6):779–808, 2000.

[11] N. Nanas, A. Roeck, and M. Vavalis. What happened to content-based information filtering? In *ICTIR '09*, pages 249–256. Springer-Verlag, 2009.

[12] S. E. Robertson and I. Soboroff. The TREC 2002 filtering track report. In *TREC*, 2002.

[13] Y. Zhang and J. Callan. Maximum likelihood estimation for filtering thresholds. In *SIGIR '01*, pages 294–302, 2001.

[14] Desmond Elliott. An empirical analysis of information filtering methods. MSc(R) thesis, 2011, University of Glasgow.

Practical Considerations when Filtering Documents

Desmond Elliott
Leif Azzopardi

It's hard to configure and tune a filtering system to ensure the user has a good experience

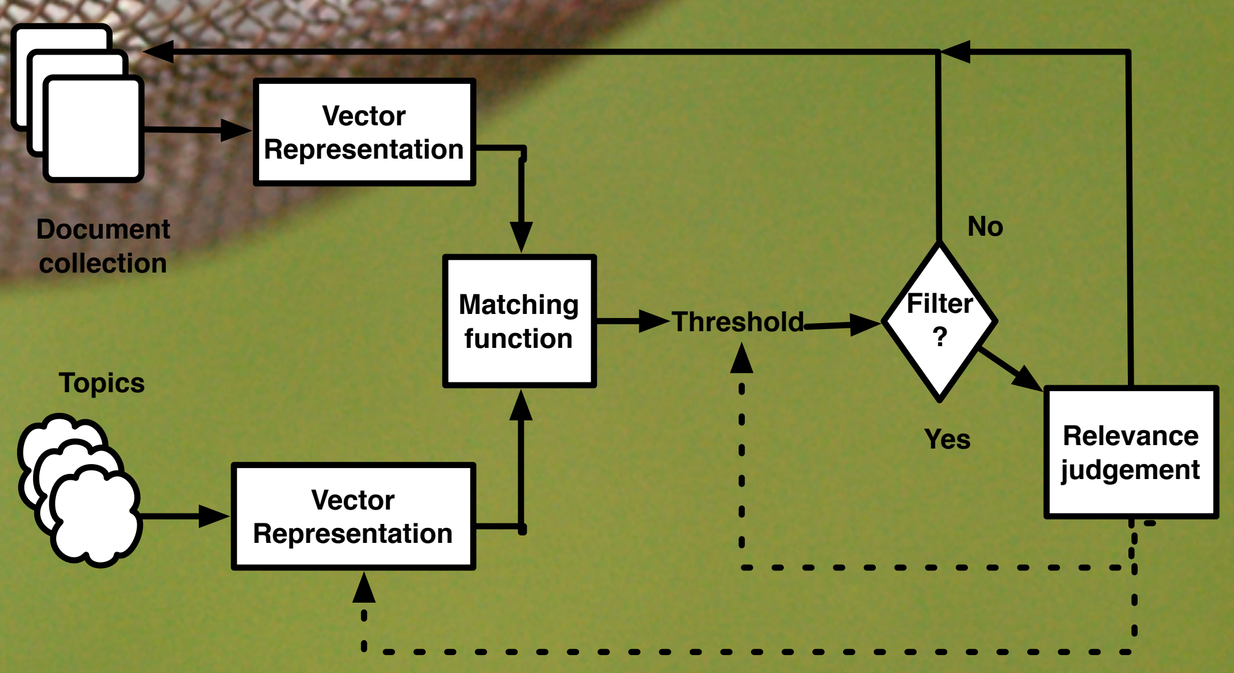
Implementing, configuring, and running an information filtering system in a practical setting is a difficult and challenging problem.

This is due to variety and configuration of available system components along with additional factors such as topic length, feedback, and system training.

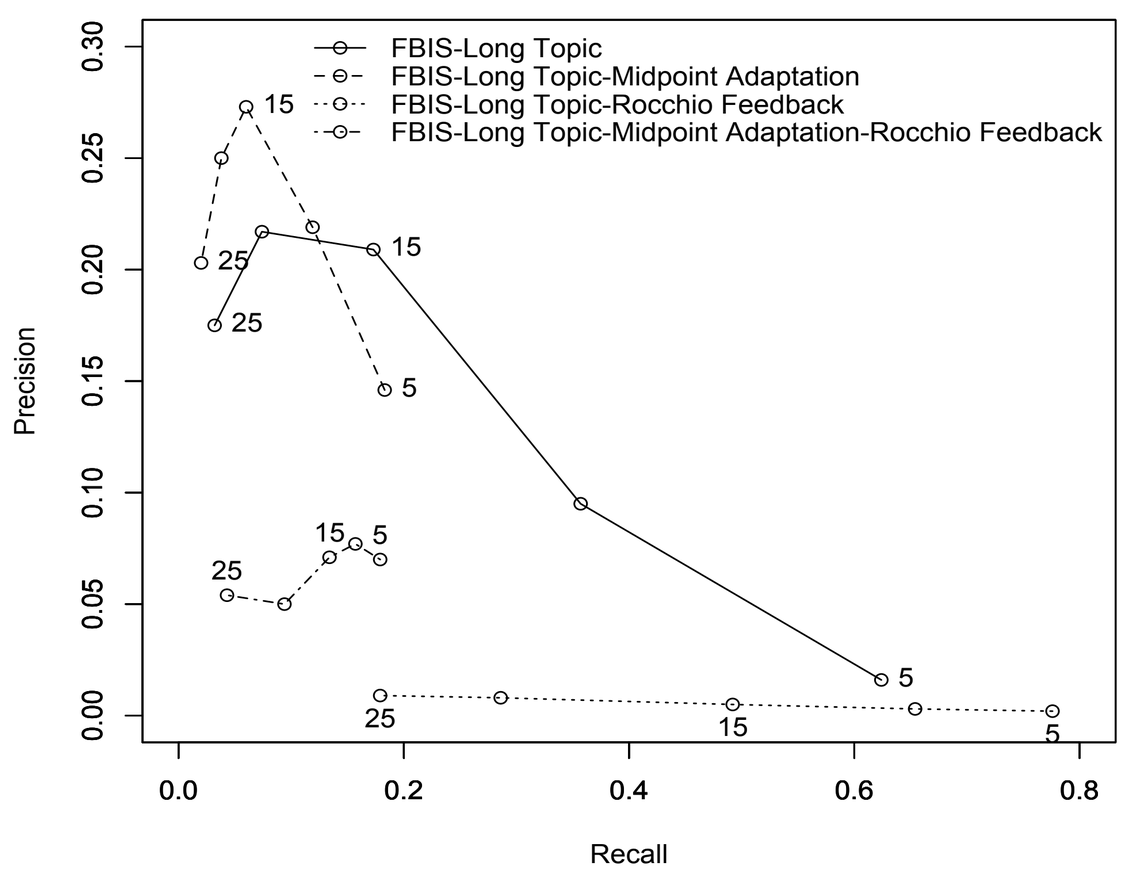
Moreover, the interplay between the different components and additional factors can lead to degraded system performance when adding or manipulating particular components.

We explore the interactions and effects of different components and some of the factors with respect to performance.

The main contribution of this work is a better understanding of how to configure filtering systems along with the possible pitfalls of applying conflicting components which harm performance and result in a poor user experience.



The effect of threshold and topic adaptation on Precision and Recall



The labels show the different initial thresholds

Guidelines for Configuration

Varying Threshold

As the threshold increases ⬆, precision increases ⬆ and recall decreases ⬇. This is observed until a point of maximum precision, after which both precision decreases ⬇ and recall continues ⬇ to decrease. With larger increases using the bounded function.

Topic Length

As the initial topic length increases, ⬆ precision increases ⬆ and recall increases. With larger increases observed when using the bounded function.

Midpoint Threshold Adaptation

As the threshold increases ⬆, precision increases ⬆, but recall decreases ⬇.

Rocchio Topic Adaptation

With further topic adaption increases ⬆ (and subsequent increases ⬆ to topic length) precision decreases ⬇ and recall increases ⬆.

Threshold and Topic Adaptation

As the threshold increases ⬆ and topic length increases, precision decreases ⬇ and recall decreases. ⬇

THE UNIVERSITY of EDINBURGH informatics

University of Glasgow | School of Computing Science

Credit: http://www.flickr.com/photos/verbeeldingskr8/3638834128/