# Beyond TREC's Filtering Track

**Nikolaos Nanas*, Victoria Uren*, Anne de Roeck†, John Domingue***

*Knowledge Media Institute,
The Open University, Milton Keynes MK7 6AA, U.K.
{n.nanas, v.s.uren, j.b.domingue}@open.ac.uk

†Computing Department,
The Open University, Milton Keynes MK7 6AA, U.K.
a.deroeck@open.ac.uk

**Abstract**

Following the withdrawal of the filtering track from the latest TREC conferences, there is a niche for new evaluation standards. Towards this end, we suggest, based on variations of TREC's routing subtask, two new evaluation methodologies. The first can be used for evaluating single, multi-topic profiles and the second for testing the ability of a multi-topic profile to adapt to both modest variations and radical drifts in user interests.

## 1. Introduction

Information Filtering (IF) systems seek to provide a user with relevant information based on a tailored representation of the user's interests, a *user profile*. The user interests are considered to be long-term. Consequently, a user may be interested in more than one topic in parallel. Also, changes in user interests are inevitable and can be both modest and radical. In addition to fluctuations in the level of interest in certain topics, new topics may emerge and interest in existing topics may be lost. Ideally, a user profile should be able to represent multiple topics of interest and their interrelations and to adapt to a variety of changes in them over time.

But typically, IF research inherits profile representations that ignore term dependencies from Information Retrieval (IR) and Text Categorisation (TC). This kind of profile can only effectively represent one topic of interest. Several single-topic profiles are required to represent a user's multiple interests. Each profile is usually adapted separately using learning algorithms that assume a steady change of interests.

The filtering track of the Text REtrieval Conference (TREC), the most serious attempt to standardise the evaluation of IF systems, reflects these practices. It only considers the evaluation of single-topic profiles and does not test the ability of IF systems to adapt to radical changes in a user's interests. Furthermore, according to TREC, IF focuses only on dynamic information sources. This is a limiting view of IF that implies assumptions which complicate the evaluation task unnecessarily.

Our research has focused on the development of an adaptive document filtering system that we call Nootropia[1]. With Nootropia, we achieved adaptive, multi-topic IF with a single user profile. TREC's methodology is not adequate for evaluating this innovative approach to IF. For that purpose, we present in this paper two alternative evaluation methodologies based on variants of TREC's routing subtask. The first can be used for evaluating single, multi-topic profiles and the second for the evaluation of a multi-topic profile's ability to adapt to both modest and radical interest changes. Following the withdrawal of the filtering track from the last two TREC conferences (TREC-12 and TREC-13) and the resulting niche in evaluation standards for IF, this is a first step towards a new standard that may accommodate further development in the field.

[1]Greek word for: "an individual's or a group's particular way of thinking, someone's characteristics of intellect and perception"

## 2. Current IF Practices

Traditionally, IF systems inherit profile representations that ignore term dependencies from research in Information Retrieval (IR) and Text Categorisation (TC). These include the dominant vector space model (Salton and McGill, 1983), probabilistic IR models (Robertson and Sparck Jones, 1976), and linear classifiers like naive Bayes, decision trees, nearest-neighbour classification and others (Sebastiani, 2002). Even in the case of connectionist approaches to IR, like neural networks (Wilkinson and Hingston, 1991) and semantic networks (Crestani, 1997), links between terms are ignored. Such linear representations can only effectively estimate the relevance of a document to a single topic of interest.

To represent multiple interests, IF systems need to maintain several single-topic profiles. A separate profile is usually built for each topic of interest based on documents that the user has pre-classified according to these topics (Amati et al., 1997). Alternatively, online clustering algorithms can be employed to incrementally identify document classes (Lang, 1995). Finally, evolutionary approaches maintain a population of linear profiles that collectively represent the user interests (Moukas and Maes, 1998).

Based on user feedback, each profile is typically adapted separately using linear learning algorithms like Rocchio's (Rocchio, 1971). These assume a steady change of interests, reflected by a constant learning coefficient (Schapire et al., 1998). Dual profiles with separate learning coefficients have also been suggested (Billsus and Pazzani, 1999). Genetic algorithms are used in the case of evolutionary IF systemss (Moukas and Maes, 1998).

These practices are not well suited to a user's multiple and changing interest. Users are required to maintain several profiles. The topics of interest are assumed to be independent. Neither their relative importance nor their topic-subtopic relations are represented. Practically, they imply a large number of parameters (e.g. learning coefficients, number of profile terms and relative importance weights) that require optimization, which may have to be performed separately for each individual user. A more user friendly and efficient solution can be pursued through a profile that can effectively represent a user's multiple interests and adapt to a variety of changes in them.

## 3. The TREC Filtering Track

TREC's filtering track reflects the above IF practices. It is only concerned with the evaluation of single-topic profiles. Furthermore, it only tests the ability of systems to adapt to modest and loosely controlled changes in the content of documents about a specific topic.

Since its start in 1992 the annual TREC conference aims to provide a standard infrastructure for the large-scale evaluation of IR systems. Its filtering track tackles the evaluation of IF systems based, since 2001, on the Reuters Corpus Volume 1 (RCV1), an archive of 806,791 English language news stories[2]. The stories have been manually categorised according to topic, region, and industry sector (Rose et al., 2002; Lewis et al., 2003). The TREC-10 filtering track is based on 84 out of 103 RCV1 topic categories. Furthermore, it divides RCV1 into 23,864 training stories and a test set comprising the rest of the stories. A recognised drawback of this split has been the large number of test documents per topic, which does not reflect a realistic IF problem (Robertson and Soboroff, 2001). TREC-11, the last TREC that included the filtering track, uses a different set of 100 topics. Fifty of them were constructed using assessors judgments on documents retrieved for a specific topic definition, by multiple retrieval and classification systems. The remaining fifty topics were constructed as intersections of pairs of RCV1 topics. This was a cost-effective way of constructing test topics out of a collection's categories. In both cases, the topics correspond to a smaller number of test documents than RCV1 topic categories (Soboroff and Robertson, 2003).

The filtering track is further divided into three subtasks: routing, batch filtering and adaptive filtering. For all subtasks, a separate, single-topic profile is built for each topic. For the routing and batch filtering subtasks, the complete training set is available for profile initialisation. In contrast, only two (TREC-10), or three (TREC-11) training documents per topic are allowed in the case of the adaptive filtering subtask. Finally, all three subtasks allow the use of any non-relevance related information from the training set.

Constructed profiles are tested against the complete test set. The output of the routing task is a ranked list of the 1000 best scoring documents. Systems are evaluated by calculating the Average Uninterpolated Precision (AUP) of this list. The AUP is defined as the sum of the precision value at each point in the list where a relevant document appears, divided by the total number of relevant documents. In the adaptive and batch filtering tasks, on the other hand, systems have to evaluate test documents in their chronological order and select a subset of them. This implies the use of thresholding for making the binary decision between selecting or discarding each document. Systems are evaluated by calculating the Utility and F-beta measure of the unordered output set.

[2]http://about.reuters.com/researchandstandards/corpus/index.asp

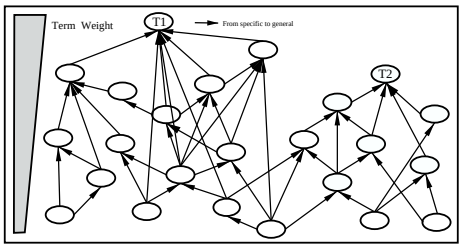


Figure 1: Hierarchical profile representing two overlapping topics of interest.

While for the batch filtering subtask the initial profile and threshold remain constant, the adaptive filtering subtask tests the ability of systems to learn a topic online and to adapt to changes in the content of the topic's test documents. Each selected document is immediately judged for relevance, and this information can be used by the system to adaptively update the filtering profile and/or adjust the threshold[3].

IF is viewed as a specialisation of IR or TC, that focuses on dynamic information sources, where the value of documents decays rapidly with time. It is assumed that a potentially relevant document has to be presented immediately to the user and that this is the only available information for profile adaptation (Robertson and Soboroff, 2001; Robertson and Soboroff, 2002). Hence, a system's performance depends a lot on threshold setting and calibration.

But while thresholding complicates the evaluation task, it is not an intrinsic part of IF. Dynamic information sources are only one of the possible application domains of IF. A user profile can be used for content-based evaluation of documents obtained from a variety of information sources (e.g. email). Furthermore, nothing prohibits the user from providing feedback to a document different to those presented by the IF system.

## 4. Adaptive Multi-Topic IF with Nootropia

In Nootropia, we use a hierarchical term network to represent a user's multiple topics of interest. This user profile is synthesised in three steps, based on a set of documents about various topics, that the user has specified as interesting. Initially term weighting is applied to identify the most informative terms in the documents. Using a sliding window, we then identify correlations between terms and finally, we order profile terms according to decreasing weight to identify topic-subtopic relations between them. This process generates a term network that formulates a

[3]For more details on TREC see: http://trec.nist.gov

separate hierarchy for each general topic discussed in the documents (Nanas et al., 2003; Nanas et al., 2004a). Figure 1 depicts a generalised hierarchical profile constructed from a set of documents about two overlapping topics.

Document evaluation is then formulated as a spreading activation model. To evaluate a document, an initial energy is deposited with those profile terms that also appear in the document. It is then disseminated from activated terms lower in the hierarchy towards activated terms further up. This establishes non-linear document evaluation that takes into account the term dependencies and topic-subtopic relations that the hierarchy represents (Nanas et al., 2004a).

Profile adaptation is achieved in Nootropia through a process of self-organisation, comprising five interrelated steps. Term weighting is applied to extract informative terms from feedback documents. The weight of existing profile terms and links is updated, incompetent terms (and their links) are removed and new terms and links are added. The process allows Nootropia to adjust structurally to both modest and radical changes in the user interests. A new hierarchy may develop to account for an emerging topic of interest and existing hierarchies that correspond to no longer interesting topics disintegrate and are eventually purged from the profile (Nanas et al., 2004b).

In contrast to current practices, with Nootropia, we can perform adaptive, multi-topic document filtering with a single-user profile. No standard methodology exists for the evaluation of this innovative approach. For that purpose, we used two variations of TREC's routing subtask.

## 5. Evaluating Multi-Topic IF Systems

To evaluate Nootropia on a multi-topic filtering problem, we experimented with profiles trained on combinations of two and three RCV1 topics. Of course, a very large number of combinations is possible. Our experiments involved six two-topic and six three-topic combinations comprising topics of varied topical proximity and collection statistics. For most combinations we have deliberately chosen topics with a small number of test documents.

A single profile was built for each one of these combinations. The training set comprised only the first 30 training documents corresponding to each topic in a combination. This amount was considered a reasonable approximation of the number of documents that a user might actually provide for profile initialisation.

Each profile was tested against the test set and evaluated on the basis of an ordered list of the best 3000 scoring documents, using the AUP measure. A separate AUP score was computed for each topic in a combination. We have increased the number of evaluated documents from 1000 (according to TREC) to 3000 for two reasons. Firstly, as an additional remedy to the large number of test documents per RCV1 topic. Secondly, the best 1000 documents can be easily dominated by the topic with the largest number of test documents, or with the strongest profile representation.

Using this methodology we conducted a series of comparative experiments between Nootropia and a traditional vector space profile representation, which produced positive results (Nanas et al., 2004a). The methodology can in general be applied for evaluating the initial performance of a multi-topic IF system. Satisfactory initial performance is necessary for engaging the user's further involvement.

## 6. Evaluating Adaptation to a Variety of Interest Changes

To evaluate Nootropia's ability to adapt to a variety of interest changes, we synthesised virtual users and simulated changes in their interests using combinations of RCV1 topics. The methodology is based on a further variant of TREC's routing subtask. We assume that changes in user's interests are reflected by variations in the distribution of feedback documents about various topics and that user feedback is not constrained to already filtered documents.

A virtual user's current interests can be defined as a combination $T1/T2/T3$ of RCV1 topics (Widyantoro et al., 2000). A radical change of interest may then be simulated by removing or adding a topic to this combination. For example, if the user in no longer interested in topic $T3$, then we may formulate this change as $T1/T2/T3 \rightarrow T1/T2$.

In this way, we have defined four learning tasks. $T1/T2$: the user is interested in two topics in parallel. $T1/T2 \rightarrow T1/T2/T3$: a new topic of interest ($T3$) emerges. $T1/T2/T3 \rightarrow T1/T2$: interest in topic $T3$ is lost. $T1/T2/T3 \rightarrow T1/T2/\neg T3$: the user explicitly specifies through negative feedback that topic $T3$ is no longer interesting (denoted with "$\neg$"). The first of the tasks does not simulate a radical change of interest.

Each topic combination in a task corresponds to a training phase, a period during which the topics of interest remain the same. During a training phase, a profile is trained online using a set of documents comprising the first 30 training documents per topic in the combination. This was done to enable a common experimental setting for all topics, including those with a small number of training documents. It implies however, that training documents are reused in both training phases of a task. Although this practice is not realistic, nevertheless, it is not statistically incorrect. Documents corresponding to a negated topic $\neg T3$ have been used as negative feedback. The training documents have been ordered according to publication date. Hence, the training set is not homogeneous, but rather reflects temporal variations in the publication date of documents about each topic. It reflects in that sense fast, but modest, fluctuations in the virtual user's interests.

To evaluate a profile, it is tested periodically during the last training phase in each task. In other words, after a radical change of interest has occurred. Every five training documents the profile is used to filter the complete test set. It is then evaluated on the basis of an ordered list of the best 3000 scoring documents, using the AUP measure. A separate AUP score was calculated for each topic.

Using this methodology we have performed a series of experiments with specific task formulations that reuse the topic combinations of the experiments described in the previous section. The results indicate Nootropia's ability to adapt to both modest variations and radical changes in a virtual user's interests (Nanas et al., 2004b).

The methodology casts the evaluation of adaptive IF systems as a routing task. It ignores the need for thresholding and concentrates instead on the ability of a single profile to evaluate documents according to user's multiple interests and to adapt to a variety of changes in them. This simplifies the evaluation task significantly. Systems return a ranked list of documents and can be evaluated using precision, recall and related measures, like AUP. Still, a combination of modest and radical changes of interest can be simulated in a controlled and easy to reproduce fashion.

## 7. Summary and Further Work

The filtering track has been removed from the last two TREC conferences. There is clearly space for improvement in the way IF systems are being evaluated. Towards this end we have questioned the basic assumptions underlying TREC's methodology. TREC concentrates on the application of IF to dynamical information sources. This implies the use of thresholding which complicates the evaluation task unnecessarily. Furthermore, and in accordance with traditional IF practices, TREC concentrates on the evaluation of single-topic profiles and of their ability to adapt to modest variations in content. But in IF, a user may be interested in more than one topic in parallel and radical interest changes are possible.

With our experimental IF system, Nootropia, we have shown that multi-topic document filtering and profile adaptation to a variety of interest changes, can be achieved with a single user profile. No standard methodology exists for the evaluation of this innovative approach. For that purpose we used two variants of TREC's routing subtask. The first evaluates the initial performance of single, multi-topic profiles and the second uses virtual users to test the ability of a multi-topic profile to adapt to both modest and radical interest changes.

The methodology is simple and well controlled. It is a first step towards a new evaluation standard for IF research. Of course, this requires a broad concensus and significant improvements. Further effort should be put in defining test topics and combinations of them and in compiling learning tasks. For now, it is important to stress that adaptive IF with a single, multi-topic profile is possible. The next evaluation standard for IF should not ignore this innovation.

## 8. References


Amati, G., D. D' Aloisi, V. Giannini, and F. Ubaldini, 1997. A framework for filtering news and managing distributed data. *Journal of Universal Computer Science*, 3(8):1007–1021.

Billsus, D. and M. Pazzani, 1999. A hybrid user model for news story classification. In *7th International Conference on User Modeling*. Banff, Canada.

Crestani, F., 1997. Application of spreading activation techniques in information retrieval. *Artificial Intelligence Review*, 11(6):453–482.

Lang, K., 1995. NewsWeeder: Learning to filter netnews. In *12th International Conference on Machine Learning (ICML95)*.

Lewis, David D., Y. Yang, T. Rose, and Fan Li, 2003. Rcv1: A new benchmark collection for text categorization research. *Journal of Marchine Learning Research*.

Moukas, A. and P. Maes, 1998. Amalthaea: An evolving multi-agent information filtering and discovery system for the www. *Autonomous Agents and Multi-Agent Systems.*, 1(1):59–88.

Nanas, N., V. Uren, A. De Roeck, and J. Domingue, 2003. Building and applying a concept hierarchy representation of a user profile. In *26th Annual International ACM SIGIR International Conference on Research and Development in Information Retrieval*. ACM press.

Nanas, N., V. Uren, A. De Roeck, and J. Domingue, 2004a. Multi-topic information filtering with a single user profile. In *3rd Hellenic Conference on Artificial Intelligence*.

Nanas, N., V. Uren, A. De Roeck, and J. Domingue, 2004b. Nootropia: a self-organising agent for adaptive information filtering. Technical Report kmi-tr-138, Knowledge Media Institute. http://www.kmi.open.ac.uk/people/nanas/kmi-tr-138.pdf.

Robertson, S. and I. Soboroff, 2001. The TREC 2001 filtering track report. In *TREC-10*.

Robertson, S. and I. Soboroff, 2002. The TREC 2002 filtering track report. In *TREC-11*.

Robertson, S. E. and K. Sparck Jones, 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146.

Rocchio, J., 1971. *Relevance Feedback in Information Retrieval*, chapter 14. Prentice-Hall Inc., pages 313–323.

Rose, T., M. Stevenson, and M. Whitehead, 2002. The Reuters Corpus Volume 1 - from yesterday's news to tomorrow's language resources. In *3rd International Conference on Language Resources and Evaluation*.

Salton, G. and M. J. McGill, 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill Inc.

Schapire, R., Y. Singer, and A. Singhal, 1998. Boosting and Rocchio applied to text filtering. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1).

Soboroff, I. and S. Robertson, 2003. Building a filtering test collection for trec 2002. In *26th Annual International ACM SIGIR International Conference on Research and Development in Information Retrieval*. ACM press.

Widyantoro, D. H., T. R. Loerger, and J. Yen, 2000. Learning user interests dynamics with a three-descriptor representation. *JASIS*.

Wilkinson, R. and P. Hingston, 1991. Using the cosine measure in a neural network for document retrieval. In *14th Annual Internation ACM SIGIR conference on Research and Development in Information Retrieval*. ACM Press.