A Comparative Study of Term Weighting Methods for Information Filtering

Nikolaos Nanas
The Open University
Knowledge Media Institute
Milton Keynes, U.K.
n.nanas@open.ac.uk

Victoria Uren
The Open University
Knowledge Media Institute
Milton Keynes, U.K.
v.s.uren@open.ac.uk

Anne De Roeck
The Open University
Department of Computing and Mathematics
Milton Keynes, U.K.
a.deroeck@open.ac.uk

ABSTRACT
The users of an information filtering system can only be expected to provide a small amount of information to initialize their user profile. Therefore, term weighting methods for information filtering have somewhat different requirements to those for information retrieval and text categorization. We present a comparative evaluation of term weighting methods, including one novel method, relative document frequency, designed specifically for information filtering. The best weighting methods appear to be those which balance exploiting user input and data from the collection.

Categories and Subject Descriptors
H.3.1. [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms
Experimentation, Measurement, Performance

Keywords
Information Filtering, Term Weighting

1. INTRODUCTION
It is generally acknowledged, that it is extremely difficult for people to find interesting information within the ever increasing volume available. To solve this problem, often referred to as *Information Overload*, mechanisms are needed for assessing the relevance of information to user interests. Research in Information Filtering (IF) has usually tackled this problem through the tailored representation of the user interests, a *user profile*.

Many approaches to user profile representation have been proposed. Typically, these approaches use weighted terms as the building blocks of the user profile representation. In this case, term weighting methods are necessary for weighting and thus selecting the most competent terms for populating the profile. Most existing term weighting methods have been introduced and evaluated in the context of Information Retrieval (IR) and Text Categorization (TC). Although a lot of these methods are applicable to the above problem, their adequacy has not yet been evaluated. IF approaches, usually adopt a term weighting method on the basis of its successful application in IR and TC. However, the IF problem has specific characteristics that distinguish it from IR and TC and that can influence existing methods in a negative way.

In this paper we present an "in-vitro" evaluation of existing statistical term weighting methods as well as introduce a novel term weighting approach. By "in-vitro" we mean that the evaluation is not bound to some specific IF approach but instead we adopt two simple document evaluation functions in order to produce general results. The evaluation takes into account the distinguishing characteristics of the IF problem. More specifically, in the next section we present a number of existing term weighting methods from both IR and TC. In section 3 we discuss those IF characteristics that can influence the selection of existing term weighting methods. We then go on to introduce a novel term weighting method that takes these characteristics into account (sec. 4). The methods are evaluated in section 5 using a slight modification of TREC's routing subtask. Our goal was to comply with existing standards as much as possible while at the same reflecting the characteristics of IF. Finally, section 6 summarizes the derived conclusions and points to future research.

2. TERM WEIGHTING METHODS
Statistical term weighting methods assume that a term's statistical behavior within individual documents (or appropriate sets of documents) reflects the term's ability to represent a document's content and/or distinguish it from other documents. A representation of a document's content that reflects accurately and in depth its various topics is characterized *exhaustive*. On the other hand, *term specificity* relates to the level of detail at which a given topic is represented by an individual term [14, 15]. A term that is specific to a topic can distinguish relevant documents in a collection. Therefore, specific terms are of particular importance when building a user profile.

Table 1: Contingency table

		Documents: Relevant	Non-relevant	Collection
Term	+	r (A)	$n-r$ (B)	n
	-	$R-r$ (C)	$N-R-n+r$ (D)	$N-n$
		R	$N-R$	N

where

- r is the number of user-specified documents that contain the term
- n is the total number of documents in the collection that contain the term
- R is the total number of user specified documents
- N is the number of documents in the collection

To assess the specificity of terms to a topic of interest to the user, term weighting methods can take into account variations in the distribution of terms within user-specified documents about that topic and within the complete document collection. A term's distribution can be expressed in terms of a contingency table (table 1) [16]. In the following paragraphs we use the table's notation to present a number of term weighting methods that have been introduced in the context of IR and TC.

2.1 Term Weighting Methods from IR

In contrast to IF, IR is traditionally concerned with satisfying an one-time information need. As a result the user has to be able to express what she is looking for in a straightforward and quick way, e.g. as a query of free language terms. The query is initially the only available information about what the user is looking for. Due to the lack of any additional relevance information, in IR, term weighting has been mainly concerned with the representation of the document space (*automatic indexing*). The corresponding methods do not usually take into account relevance information. They are based only on term statistics in the complete document collection in order to appropriately weight the index terms, i.e. the terms used to describe the documents. Therefore, these methods cannot explicitly measure the specificity of terms regarding the topics of interest to the user. Nevertheless, when profile terms are selected out of the unique terms in the user specified documents, relevance information is implicitly taken into account. We focus on term weighting methods used for automatic indexing that can be applied for selecting those unique terms in the user-specified documents that are more specific in the document collection. A more extensive survey of term weighting methods used in automatic indexing can be found in [6].

Weighting of query terms can be accomplished if additional relevance information is available. In IR, such information can be acquired by user feedback to the documents retrieved so far. This implies a broader information seeking episode comprising more than one query about the same information interest. Methods that take advantage of any additional relevance information to weight query terms can be similarly applied to weight the unique terms in the user specified documents.

2.1.1 Inverse Document Frequency (IDF)

IDF is one of the most widely used term weighting methods for estimating the specificity of terms in a document collection. It is based on the idea that if a term appears in only a few of the documents in the collection then such a term is expected to be a good discriminator of these documents. Sparck Jones [15], proposed equation 1 as a way of calculating the IDF weight of a term t and showed that its usage significantly improves retrieval performance compared to unweighted retrieval.

$$IDF_t = log_2 \frac{N}{n} \tag{1}$$

2.1.2 Residual Inverse Document Frequency (RIDF)

RIDF is a variation of IDF that assigns weights to terms according to the difference between the logs of the actual IDF and its prediction by a Poisson model [2]. According to Manning and Schütze [8], the most common way of calculating the RIDF of terms is given by equation 2, where $\lambda_t = cf_t/N$ is the average number of occurrences of term t per document and $1 - p(0; \lambda_t)$ is the Poisson probability that t appears at least once in a document.

$$RIDF_t = idf_t + log_2(1 - p(0; \lambda_t)) \tag{2}$$

2.1.3 Query Term Weighting

Robertson and Sparck Jones proposed four methods for the probabilistic weighting of search terms, based on the *binary independence retrieval model* [12]. The four methods correspond to the combinations between two independence assumptions and two ordering principles. Robertson has emphasized the difference between problems where complete relevance information is available (*retrospective*) and problems where estimations based on incomplete information are required (*predictive*). For predictive problems, he suggested a variation of the simple, retrospective version of the methods. Out of the four proposed methods we focus on the retrospective version of the first (eq. 3) and the predictive version of the fourth (eq. 4). The latter was shown to be the best performing approach.

$$F1_t = log\frac{r/R}{n/N} = log\frac{r}{R} - log\frac{n}{N} \tag{3}$$

$$F4_t = log\frac{(r+0.50)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \tag{4}$$

2.2 Term Weighting Methods from TC

TC is concerned with the automatic classification of documents according to relatively static topic categories. As a consequence, extensive and accurate relevance information can be made available. In categorising magazine articles for example, there is often a large document collection (training collection) that has been manually pre-classified according to a number of predefined topics [9]. For each one of the topics of interest thousands of relevant documents may be available. In addition, the classification is usually performed by more than one person and the topic categories are coarse enough to facilitate the classification of thousands of documents. Therefore, we can, with some confidence, treat any documents that have not been assigned to a specific topic as non-relevant to that topic.

The existence of extensive and accurate relevance information allows the application of machine learning algorithms to the classification task, using, for instance, *decision trees* and *neural networks* [1, 7]. However, most of these machine learning algorithms cannot easily cope with the high dimensionality of the native feature space, i.e. the complete set of unique terms in the training collection. Term weighting is used to reduce the feature space to those terms that are more specific to the topics of interest.

IF can be approached, as a binary classification problem. According to this view the goal of IF systems is to classify documents as either relevant or non-relevant to the user [11]. Therefore, term weighting methods that have been introduced in the context of TC can be used for selecting those terms in the user specified documents that are more specific to the user interests. In the following paragraphs we present the binary version of a number of well established term weighting methods from TC. Most of the presented methods have been evaluated on an m-ary text classification problem by [18].

2.2.1 Relevant Document Frequency (RDF)

RDF can be considered as the binary version of document frequency. Document frequency has proved successful in m-ary classification problems, because it identifies terms that occur across many topics [18]. Its application to a binary classification problem would however result in frequent, non-specific terms being selected. RDF on the other hand, exploits relevance information to identify terms that occur frequently within the documents of interest (eq. 5). The assumption here is that those terms are more specific to the documents' topic than terms that occur less.

$$RDF_t = r \tag{5}$$

2.2.2 Information Gain (IG)

IG is an information-theoretic metric that measures the difference in the entropy of category prediction by knowing the presence or absence of a term in a document. Equation 6 is the applicable binary version of the metric.

$$\begin{aligned} IG_t &= -Pr(rel)logPr(rel) + Pr(t)Pr(rel|t)logPr(rel|t) + \\ &+ PR(t)Pr(rel|t)logPr(rel|t) = \\ &-\frac{R}{N}\cdot log\frac{R}{N} + \frac{r}{N}\cdot log\frac{r}{n} + \frac{R-r}{N}\cdot log\frac{R-r}{N-n} \end{aligned} \tag{6}$$

2.2.3 Mutual Information (MI)

MI is another measure derived from information theory. It gauges the reduction in uncertainty of one random variable when we know about another. The metric is commonly applied for identifying term collocations. In a similar way it can be used for measuring the association between a term and a specific topic of interest (eq. 7). We should note that the equation is identical to F1 (eq. 3).

$$F1/MI_t \approx log\frac{A\times N}{(A+C)\times(A+B)} = log\frac{r/R}{n/N} \tag{7}$$

2.2.4 χ^2 chi square (CHI)

CHI is similar to MI in that it measures the lack of independence between two variables. It calculates the difference between the observed frequencies in the contingency table and the frequencies expected under the independence assumption. If the difference is large, then we can treat the variables as not independent. For the problem at hand, χ^2 is applied to measure the lack of independence between a term and the user-specified topic of interest (eq. 8).

$$\begin{aligned} \chi^2 &= \frac{N\cdot(AD-CB)^2}{(A+C)\cdot(B+D)\cdot(A+B)\cdot(C+D)} = \\ &= \frac{N\cdot(rN-nR)^2}{R\cdot n\cdot(N-R)\cdot(N-n)} \end{aligned} \tag{8}$$

3. THE CHARACTERISTICS OF USER-SPECIFIED INFORMATION

When building a user profile, the user is of course the only source of information about what is of interest. Foltz and Dumais have shown that in IF, it is more effective and easier for users to express their interests in terms of a small set of documents than as lists of terms and/or phrases [4]. A set of user-specified documents about a topic of interest provides both the pool of candidate profile terms and the necessary information for their weighting and selection. The set can either be specified explicitly as part of the profile initialization process [17], or implicitly based on the user feedback to already filtered documents.

In the case of profile initialization and in contrast to TC, the user neither has the time nor the inclination to specify a large number of documents for each one of her topics of interest. However, a small set of relevant documents can be easily compiled from either saved documents or the user's bookmarks. Furthermore, while in TC the topic categories of interest are known in advance, in IF we can neither predefine the number of topics that the user is going to specify nor their topical proximity. As a consequence, we cannot confidently treat documents that are not assigned to a topic as non-relevant. It is therefore preferable to treat each set of documents individually and not as part of a preclassified collection. Nevertheless, even a small number of documents about an interesting topic provides more information and a larger choice of terms than a query. Term weighting methods can be applied to weight and select those unique terms in the set of documents that are more specific to its underlying topic.

User feedback on documents that have already been filtered by the profile provides additional relevant/less relevant documents. In this case, term weighting can be applied to weight the terms in these document sets. The weighted terms can then be used to either update the weights of profile terms or for adding new terms in the profile.

In addition to the user specified documents, term weighting methods can take into account information extracted from the collection. Collection statistics identify frequent terms which have low discrimination power. Traditionally in IF, the collection is considered dynamic, but in a lot of practical applications a collection does exist. An example is the documents stored in a company's repository. Even if a collection does not exist, term statistics in general language can be used[1].

[1] As part of the UC Berkeley and Stanford University Digital

In conclusion, term weighting methods that are appropriate for weighting and selecting the profile terms, should take advantage of the information provided by a small set of user specified documents. The additional information that is acquired from the collection can also be important.

4. A TERM WEIGHTING METHOD FOR INFORMATION FILTERING

In this section we introduce a novel term weighting approach that complies with the above requirements. *Relative document frequency* (RelDF) is a measure of the relative importance of terms within the user specified documents and a general collection of documents. The essence behind the approach is analogous to the *relative frequency* technique that has been suggested by Edmundson and Wyllys [3] (hence the adopted name). Based on the assumption that special or technical words are more rare in general usage than in documents about the corresponding subjects, they presented a number of ways for assessing the relative frequency of terms within a document and a general collection.

In a similar way, we assume that terms pertaining to the topic of interest to the user will appear in a larger percentage of the user specified documents than in the general collection. The method assigns to each term, a weight in the interval (-1,1), according to the difference between the term's probabilities of appearance in the user specified documents and in the general collection (eq. 9). While the first part of the equation ($\frac{r}{R}$) favors those terms that exhaustively describe the user specified documents and therefore the underlying topic of interest, the second part ($-\frac{n}{N}$) biases the weighting towards terms that are specific within the general collection.

$$RelDF = \frac{r}{R} - \frac{n}{N} \tag{9}$$

5. EVALUATION

We evaluated the term weighting methods using a slight variation of the TREC-2001 routing subtask. Our goal was to comply with an existing and well established evaluation methodology as much as possible, while at the same time to take into account the small number of user-specified documents. Our intention is not the proposal of a new evaluation standard. Nevertheless, the evaluation procedure can be easily reproduced by other researchers.

5.1 Experimental Setup

The Text REtrieval Conference (TREC) has been held annually since 1992 and its purpose is to provide a standard infrastructure for the large-scale evaluation of IR systems. TREC-2001 adopts the Reuters Corpus Volume 1 (RCV1). The latter is an archive of 806,791 English language news stories that recently has been made freely available for research purposes[2]. The stories have been manually categorized according to topic, region, and industry sector [13]. The TREC-2001 filtering track is based on 84 out of the 103 RCV1 topic categories. Furthermore, it divides RCV1 into 23,864 training stories and a test set comprising the rest of the stories[3].

According to TREC's routing task systems are allowed to use the complete relevance information and any non-relevance-related information from the training set. Systems are evaluated on the basis of the best 1000 scoring documents, using the *average uninterpolated precision* (AUP) measure. 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.

We have experimented using a slight variation of TREC's routing subtask. To minimize the time needed for each experiment we have only used the first 10 out of the 84 TREC topics (R1-R10). Furthermore, we have attempted to reflect more realistically the amount of relevance information that a user can provide for each topic of interest. While according to the routing subtask systems are allowed to use the complete relevance information provided by the training set, we have performed a series of experiments using only the first 10, 20, 30 and 40 relevant documents per topic – far less than the hundreds provided for most of the topics by the training set.

Each of the 10 topics, was preprocessed by stop word removal and stemming using Porter's algorithm, to reduce the space of unique terms in the relevant documents. The remaining terms were weighted by each method and a topic specific profile was constructed using the most competent terms. In order to evaluate the effect of the number k of profile terms on the profile's filtering performance, different profiles were constructed for each $k \in \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 40, 60, 80, 100\}$. More results were produced for profiles with a small number of terms, to recognize term weighting methods that could identify the most informative terms in only a small number of extracted terms. In summary, a different topic-specific profile was constructed for each possible combination of term weighting method, topic, number of relevant documents and number of profile terms. In total 1120 profiles were evaluated for each of the term weighting methods.

The profiles were then used to assess the relevance of the documents in the test set. Independence between profile terms and binary indexing of documents were assumed. For each profile P and document D, two different evaluation functions were adopted. In the first case, documents were evaluated according to the inner product measure [5]. A document's relevance R was calculated as $R_{P,D} = \sum w_i * dw_i$, where w_i and dw_i are respectively the weights of a term t_i in the profile and in the document. Since binary indexing was assumed the previous equation can be simplified to equation 10. In that sense a document's relevance is calculated as the sum of the weights of profile terms that it contains.

$$R_{P,D} = \sum_{t \in D} w_t \tag{10}$$

We have also experimented with evaluating a document by the product of the weights of profile terms that it contains.

Library projects, the term document frequency and rank for all the terms on the Web has been made freely available at: http://elib.cs.berkeley.edu/docfreq/index.html

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

[3]For more details on the TREC 2001 filtering track see: http://trec.nist.gov/data/t10_filtering/T10filter_guide.htm

Table 2: Term weighting methods

Abbreviation for method	Abbreviation
Information Gain	IG
Relative Document Frequency	RelDF
Relevant Document Frequency	RDF
χ^2 (chi square)	CHI
Robertson's 4th Formula (predictive)	F4
Robertson's 1st Formula (retrospective) & Mutual Information	F1/MI
Inverse Document Frequency	IDF
Residual Inverse Document Frequency	RIDF

In this case a document's relevance R was calculated by equation 11. Although this approach is relatively ad-hoc, it was motivated by the joint probability of independent features. Our goal was to find another way of uniformly comparing the term weighting methods. The multiplication approach is applicable as long as term weights are greater than one. Only then does the product of term weights increase with the number of terms. Consequently, the weights of profile terms have been scaled so that no weight is less than one. The drawback of the multiplication approach is that it can overestimate the relevance of a document containing too many profile terms, even if these terms are not the most informative terms in the profile.

$$R_{P,D} = \prod_{t \in D} w_t \tag{11}$$

5.2 Results

Each profile was used to evaluate the documents in the test set. The AUPs of the profiles corresponding to each method were averaged over the different topics (R1-R10) and the different numbers of relevant documents (10, 20, 30, 40). Figures 1 and 2 respectively present the results using summation of weights and multiplication of weights. In these graphs, the X axis corresponds to the number of profile terms and the Y axis to the average AUP score. A different line has been plotted for each term weighting method. Table 2 summarizes the names of the evaluated term weighting methods and the corresponding abbreviations. Finally, in table 3 each method's score for different numbers of profile terms has been averaged to a single overall score value.

The results reveal a significant difference in the performance levels of IG, RelDF, RDF and CHI in comparison to F4, F1/MI, IDF and RIDF. In other words methods from TC appear to perform better than methods from IR. These first four methods are those biased towards the information provided by the user. They favor terms that appear in a lot of relevant documents to those appearing in only a few. In contrast, the smoothing effect of logarithm in combination with, the small number of user specified documents and the substantially larger number of documents in the collection, biases F4 and F1/MI towards information acquired from the collection (eq. 3 and 4). Large differences in the document frequency of terms are more strongly taken into account than small differences in their relevant document frequency. This negative effect of algorithmic smoothing is evident in the difference between the performance of RelDF and F1/MI. Although both methods use the same statistics, the application of logarithms results in reduced performance for F1/MI. The importance of the user specified information is also highlighted by the poor performance of IDF and RIDF that do not take into account the relevant document frequency of terms. However, the information provided by the user is not sufficient for optimum performance. Despite the fact that RDF performs substantially better than IDF and RIDF, RelDF performs even better. The difference in their performance is apparently due to the collection statistic that RelDF takes into account (second fraction of equation 9).

Apparently, the way a term weighting method combines information provided by the user and information acquired from the collection is a significant performance factor. While both kinds of information should be taken into account, what the user provides is of increased importance. This finding is not only supported by the higher overall score of the first four methods of table 3 but also by the increased performance of the first three of them for small number of profile terms. IG, RelDF and RDF have the ability to identify the most informative terms within only a small number of extracted terms. The opposite happens in the case of F4, F1/MI, IDF and RIDF. CHI appears to behave in a way intermediate to these two extremes.

Table 3 presents the evaluated methods by decreasing order of overall score. IG is the best performing approach while RelDF represents a promising alternative. In addition to its competitive performance improved variations to RelDF can be formulated. Variations of the form, $a \cdot \frac{r}{R} - b \cdot \frac{n}{N}$ or $(\frac{r}{R})^a - (\frac{n}{N})^b$, where a and b define the relative importance of the two fractions and thus of the corresponding kind of information, may result in even better performance. We are currently experimenting with $(\frac{r}{R})^2 - \frac{n}{N}$.

Despite its simplicity, the competitive performance of RDF is not surprising. RDF takes into account the important user-provided information. In addition, its results are analogous to those presented by [18], for its m-ary counterpart, document frequency. It is the performance of CHI that is relatively unexpected. CHI is the worst of the four methods from TC. This is possibly due to CHI's m-ary nature. While in an m-ary classification problem it is usually secure to treat documents not pertaining to a certain topic as non-relevant to that topic, we have already noted that in our case not all of the documents in the training set that pertain to a certain topic are used for the construction of the corresponding profile.

Out of the methods from IR, F4 is the best performing one. Its superior performance over F1 confirms the results presented by [12]. IDF and RIDF are the worst performing approaches probably because they do not exploit information supplied by the user. It is however interesting to note that RIDF performs slightly better than the rest of the IR methods for small number of extracted terms. This characteristic of RIDF can be attributed to its Poisson distribution component that takes into account the frequency of occurrence of terms in the user specified documents. As a result the user provided information influences to some extent the weighting of terms.

As expected the results using multiplication of weights for

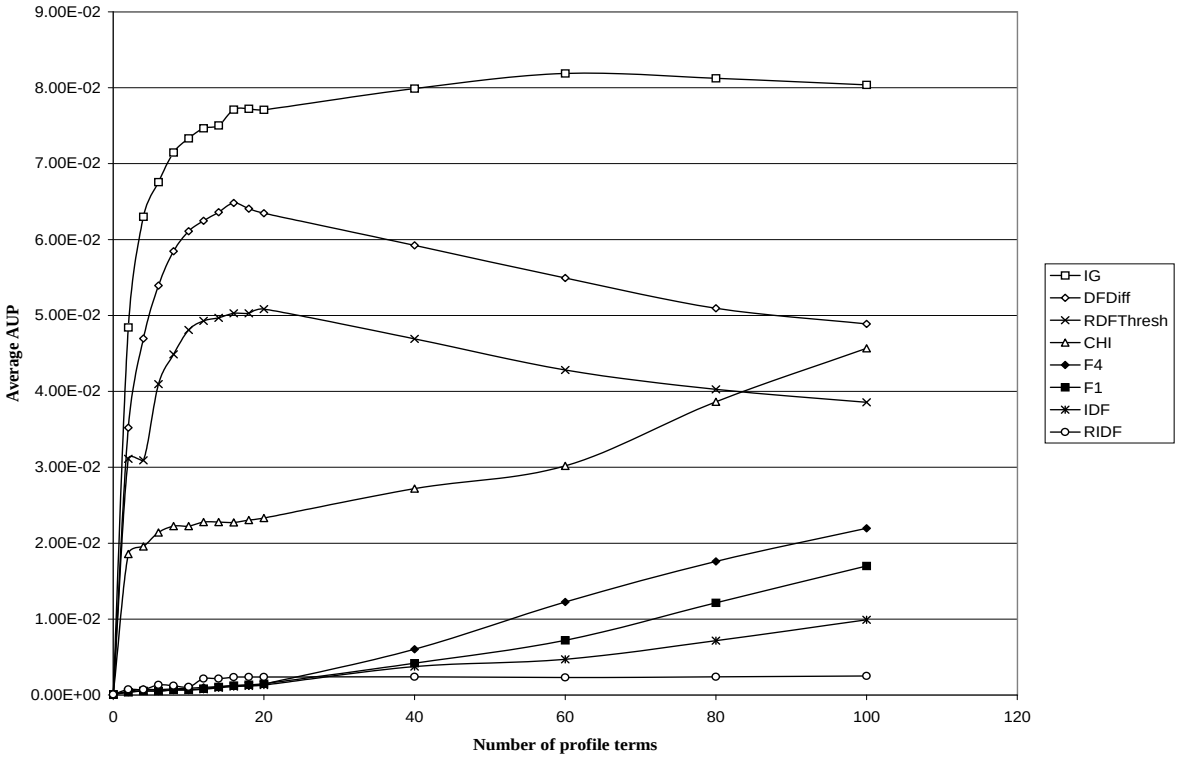

Figure 1: Results of experiment using summation of weights (eq. 10)

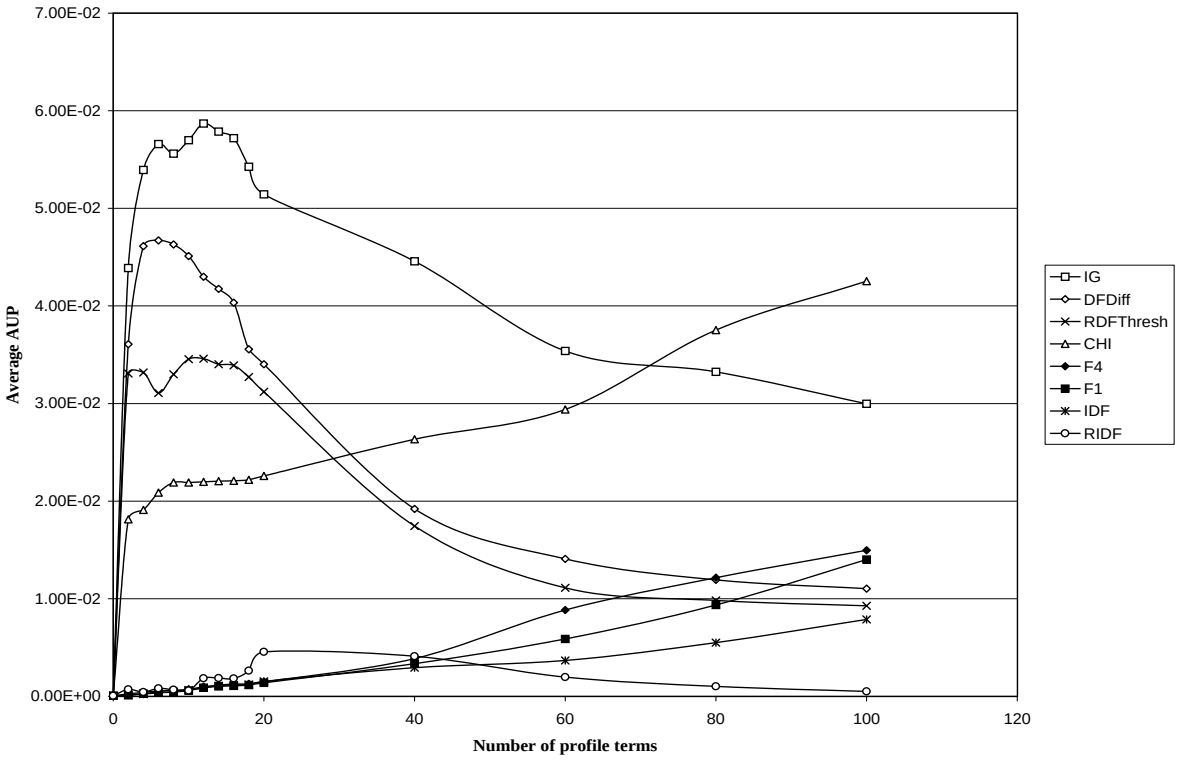

Figure 2: Results of experiment using multiplication of weights (eq. 11)

Table 3: Overall Score

	Evaluation Function	
Method	Sum (eq. 10)	Product (eq. 11)
IG	0.07346	0.04926
RelDF	0.05629	0.03366
RDF	0.04392	0.02707
CHI	0.02574	0.0249
F4	0.00482	0.00346
F1/MI	0.00352	0.00285
IDF	0.0024	0.00197
RIDF	0.00186	0.00168

document evaluation are worse than those using summation. Nevertheless, in both cases the behavior of the evaluated methods is analogous both in terms of relative performance and in terms of performance trend. Therefore, both document evaluation approaches confirm the above findings. This is logical since the assignment of weights is done irrespective of the adopted document evaluation approach. Term weights represent the specificity of terms to the user-specified documents. The incorporation of these weights into a meaningful document evaluation function is a serious research question that was not the focus of this paper.

6. CONCLUSIONS AND FUTURE RESEARCH

In this paper we have presented an evaluation of a large number of term weighting methods for the problem of identifying informative terms within a small number of user-specified documents about a topic of interest. The documents can be provided either as part of the user's profile initialization process or as a result of user feedback. The extracted terms can thus be used either to populate the initial profile or to update an existing one. Existing methods, that have been introduced in the context of IR and TC, and a novel term weighting approach (RelDF), have been evaluated on an appropriate modification of the TREC-2001 routing subtask.

The results indicate that methods from TC are more appropriate for IF than methods from IR. These methods favor information provided by the user specified documents, over information from the collection. IG is the best performing approach while RelDF appears to be a promising alternative. The results can be used as evidence for the appropriate choice of a term weighting method by systems that focus on other aspects of personalized IF. In addition the easy reproduction of the experimental setup and the basic document evaluation functions that have been adopted allow the use of the results as a baseline for comparisons with more elaborate IF approaches.

The experiments presented in this paper are part of ongoing research for the development of an adaptive information filter, called *Noo-tropia*. Noo-tropia uses a three layered concept hierarchy in order to take term dependence into account in the way user interests are represented [10]. Adaptation is achieved on the basis of an economic model. The results presented here allowed us to choose IG and RelDF for the rest of our experiments. They also provide the base-line against which we will judge improvements in IF performance.

7. ADDITIONAL AUTHORS

Additional author: John Domingue (Knowledge Media Institute, email: `j.b.domingue@open.ac.uk`).

8. REFERENCES

[1] H. Chen. Machine learning for information retrieval: Neural networks, symbolic learning and genetic algorithms. *Journal of the American Society for Information Science*, 46(3):194–216, 1995.

[2] K. W. Church. One term or two? In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 310–318, Seattle, WA USA, 1995.

[3] H. P. Edmundson and R. E. Wyllys. Automatic abstracting and indexing - survey and recommendations. *Communications of the ACM*, 4(5):226–234, 1961.

[4] P. W. Foltz and S. T. Dumais. Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM*, 35(12):51–60, 1992.

[5] W. P. Jones and G. W. Furnas. Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society of Information Science*, 38(6):420–442, May 1986.

[6] K. Kageura and B. Umino. Methods of automatic term recognition. *Terminology*, 3(2):259–290, 1996.

[7] D. D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. In *Symposium on Document Analysis and Information Retrieval*, 1994.

[8] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[9] M. Moens and J. Dumortier. Text categorization: the assignment of subject descriptors to magazine articles. *Information Processing and Management*, 36(6):841–861, 2000.

[10] N. Nanas, V. Uren, A. De Roeck, and J. Domingue. Building and applying a concept hierarchy representation of a user profile. *Submitted to the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003. http://kmi.open.ac.uk/people/nanas/nanasSigir2003B.ps.

[11] M. J. Pazzani. Representation of electronic mail filtering profiles: A user study. In *International Conference on Intelligent User Interfaces*, New Orleans, LA USA, 2000.

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

[13] T. Rose, M. Stevenson, and M. Whitehead. The reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, 2002.

[14] G. Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372, 1973.

[15] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20, 1972.

[16] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition, 1979.

[17] W. Winiwarter. Pea - a personal email assistant with evolutionary adaptation. *Internation Journal of Information Technology*, 5(1), 1999.

[18] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97)*, 1997.