Interactive Retrieval Based on Faceted Feedback

Lanbo Zhang, Yi Zhang
School of Engineering
UC Santa Cruz
Santa Cruz, CA, USA
{lanbo, yiz}@soe.ucsc.edu

ABSTRACT

Motivated by the commonly used faceted search interface in e-commerce, this paper investigates interactive relevance feedback mechanism based on faceted document metadata. In this mechanism, the system recommends a group of document facet-value pairs, and lets users select relevant ones to restrict the returned documents. We propose four facet-value pair recommendation approaches and two retrieval models that incorporate user feedback on document facets. Evaluated based on user feedback collected through Amazon Mechanical Turk, our experimental results show that the Boolean filtering approach, which is widely used in faceted search in e-commerce, doesn't work well for text document retrieval, due to the incompleteness (low recall) of metadata assignment in semi-structured text documents. Instead, a soft model performs more effectively. The faceted feedback mechanism can also be combined with document-based relevance feedback and pseudo relevance feedback to further improve the retrieval performance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

interactive retrieval, faceted feedback, relevance feedback, metadata-based retrieval

1. INTRODUCTION

A personalized search or filtering system usually suffers from the "cold start" problem, where the system performs poorly when it has little training data about new users. Researchers have proposed some approaches trying to alleviate this problem. One direction is to borrow information from other users [20, 24]. For example, the idea in [24] is to learn a prior of user interests based on the behaviors (training data) of all users, and learn the user profile for a new user based on both the prior and the training data from this user. Another direction is to develop user interaction mechanisms to collect more information from users[18]. In this paper, we focus on the second direction. We aim to study a new interactive user feedback mechanism that helps retrieval systems learn more about user information needs with limited user interactions.

Faceted search has gained great success in e-commerce domain over the past years, and most popular online retailers, such as Amazon and eBay, now provide faceted search interfaces. On faceted-search-enabled websites, buyers can narrow down the list of products by putting constraints on a group of merchandize facets, such as category, price, brand, size, etc. Well designed faceted search has been shown to be understood by the average user [11]. This motivates us to explore whether we can adapt the faceted search idea to the general purpose document retrieval. In each domain, documents have their own facets, which might be manually assigned or generated automatically. These facets are usually stored in the form of faceted document metadata. Each metadata field corresponds to a facet type, and the specific value assigned to a field for a particular document is a facet value.

Users might have preferences for certain document facets. For example, Chinese readers prefer reading news written in Chinese; some students enjoy learning by reading slides which are usually in the "ppt" format rather than reading long papers which are usually in the "pdf" format; researchers are usually interested in papers within their own subjects; movie viewers might have preferences on movie genres, directors, or casts; online buyers might have preferences on brands, colors, etc. In all these cases, users have clear ideas about some facets of their interested documents, and this information might help the system learn users' preferences and interests. Ideally, users would provide structured queries to describe their information needs more accurately. However, INEX experiments on structured documents retrieval and previous research on log analysis found that people do not use structure in their queries frequently, or use them incorrectly and thus do not improve search effectiveness if they are forced to do so [8].

In this paper, we explore a simple interactive user feedback mechanism based on document facets, called faceted feedback. In this mechanism, instead of letting users pro-

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

vide relevance feedback on documents or create structured queries actively, the system suggests faceted constraints (in the form of facet-value pairs) and users can choose interesting facet-value pairs to improve the returned documents.

We study two major problems of designing a faceted feedback based retrieval system. First, how to recommend facet-value pairs to users. In e-commerce domain, the candidates of facets and possible values for products are usually manually designed. To make it applicable in general purpose document retrieval, automatic facet recommendation is needed. In this paper, we investigate four approaches to recommending good facet-value pairs. Secondly, we study how to use user faceted feedback in retrieval. Existing e-commerce websites often use a Boolean filtering strategy while retrieving products. However, this may not be a good approach for all domains, since the document metadata is usually imperfect, and the rigid Boolean model may miss relevant documents and hurts the system recall. Thus we also propose a soft retrieval model. In this model, a document that meets a users elected faceted constraint gets a certain number of credits.

The proposed faceted feedback mechanism may have the following advantages. First, the suggested facet-value pairs are usually short and easy to understand. Compared with document-based feedback, this may reduce the cognitive overload of the user and thus is more likely to be adopted by the average user. Users can quickly select multiple facet-value pairs in a short time, so the system might get more user feedback. Second, it may help a user better understand the corpus, how the engine works, and train users in how to form better queries.

The rest of this paper is organized as follows. In section 2, we talk about the related work. Section 3 is the focus of this paper, and describes the faceted feedback mechanism. We propose four facet-value pair recommendation methods and two retrieval models in this section. In section 4, we describe the methodology of our experiments. Section 5 gives the experimental results and the corresponding analysis. Section 6 concludes this paper.

2. RELATED WORK

Many existing search engines equate user information needs with a keyword query, assuming that a user knows what words to use to best describe his or her information need. However, a user's information need is characterized by complex user criteria that are not included in a simple keyword query. Relevance feedback is a commonly used query refinement technique that can be traced back to 1960s. The basic idea is to rely on user interactions to better capture the user information need.

Document-based relevance feedback is one of the most widely used explicit feedback mechanisms. In this scenario, users are asked to provide feedback on the relevance of delivered documents. Many approaches have been proposed to incorporate document relevance feedback into retrieval. For example, Rocchio proposed to combine the original query vector with the center of relevant documents and the center of non-relevant documents [18]. Zhai et al. proposed to estimate a feedback topic model based on user feedback using Maximum Likelihood Estimation (MLE) in the language modeling approach [23]. Zhang et al. proposed to use the Bayesian logistic regression model combined with Rocchio algorithm [24]. Also, several approaches have been proposed to actively select good documents for users to provide relevance feedback. The simplest way is to choose the top ranked documents since they are most probably relevant. Others also tried other approaches, such as to choose documents with presumably good qualities (e.g., Wikipedia articles), or to choose a diversified set of documents based on document clustering or active learning [19].

A special type of document-based relevance feedback is pseudo relevance feedback. In this case, the top ranked documents are assumed to be relevant and used to modify the query based on the document feedback algorithms described above. Though the assumption is not true, pseudo relevance feedback has been proven effective in improving retrieval performances for short queries [12]. Kelly et al. found that pseudo-relevance feedback performs better for recall-oriented measures [13].

Term-based relevance feedback is to let users select relevant terms from a group of candidates suggested by query expansion techniques. However, research on term-based feedback have mixed results: some found it effectively improves retrieval performance [10, 22], while others found no obvious improvement [4].

Raghavan et al. proposed to use feedback on both instances and features and proposed a unified framework that can be used to combine document-based relevance feedback and feature-based relevance feedback [15].

Our work is motivated by early work in relevance feedback, and differs by focusing on retrieving semi-structured documents with faceted metadata. Anick et al. proposed to extract faceted terminologies automatically from the document text and let users provide relevance feedback on these faceted terminologies [5]. However, the facets in this paper refer to faceted terminologies, usually noun phrases. [9] proposed to get user feedback about controlled indexing vocabulary and got promising results on OHSUMED data set. However, existing research didn't provide detailed description about the algorithms or any quantitative evaluation with real users.

3. FACETED FEEDBACK

Unlike document-based relevance feedback mechanism which asks users to give feedback on the relevance of documents, faceted feedback allows users to give feedback on document metadata fields. In this paper, each metadata field is called a facet, and a facet (f) with a specific value (v) is called a facet-value pair ($f : v$). Each facet-value pair represents a faceted constraint on returned documents, E.g., language:Chinese, format:ppt, subject:IR, genre:comedy.

3.1 Facet-value pair recommendation

To avoid overwhelming users with many facet-value pair candidates, the system needs to recommend a small number of facet-value pairs that are most probably interesting to a user. A good recommendation approach is crucial in the faceted feedback mechanism. Intuitively, the recommended facet-value pairs should be good in two respects: 1) they have a high probability of being relevant and thus chosen by the user; 2) they maximize the learning benefits if known to be relevant. Based on the first respect, we propose four facet-value pair recommendation methods. We will investigate the second respect in our future work.

3.1.1 Top Document Frequency (TDF)

The first approach is to select the most frequent facet-value pairs occurring in the top N ranked documents returned by a baseline retrieval algorithm using the initial query. We calculate the frequency of each facet-value pair in the top N documents, which is called "Top N Document Frequency" (TDF). The top K most frequent facet-value pairs are chosen as candidates to present to the user. The underlying assumption is that the more frequently a facet-value pair appears in the top ranked documents, the more likely the user will like it.

3.1.2 TDF-IDF

In the term-based feedback literature, researchers have concerns about using the most frequent terms from the top ranked documents, because a lot of common noisy terms are likely to be selected [23]. To avoid similar problems in faceted feedback, we consider another feature of facet-value pairs: the Inverse Document Frequency (IDF), which has a similar definition to the IDF of terms. When scoring a facet-value pair, we use the product of its top N document frequency (TDF) and IDF:

$$score(f:v,q) = tdf(f:v,q,N) * idf(f:v) \tag{1}$$

where $f:v$ is a facet-value pair, q is the initial query, and $tdf(f:v,q,N)$ is the top N document frequency of $f:v$ for query q.

The motivation of using IDF is twofold: 1) a facet-value pair that appears rarely in the whole corpus while frequently in top ranked documents has a high probability to be relevant; 2) the retrieval system gets more benefits by knowing a rare facet-value pair covering a small number of documents being relevant than a frequent one.

3.1.3 Query Likelihood (QL)

Our third method is based on the language modeling approach. The query likelihood given each facet-value pair $P(q|f:v)$ is estimated. The facet-value pairs with the largest query likelihoods are chosen as the candidates.

$$P(q|f:v) = \prod_{w_j \in q} P(w_j|f:v)^{c(w_j,q)} \tag{2}$$

where $c(w_j,q)$ is the frequency of w_j in the query q, and

$$P(w_j|f:v) = \sum_{d \in \mathbf{C}} P(w_j|d)P(d|f:v) \tag{3}$$

This is a "translation" model motivated by Berger et al. [6]. $\mathbf{C}$ is the whole corpus, $P(w_j|d)$ is the language model of document d, and $P(d|f:v)$ is assumed to be uniform over all documents that contain $f:v$.

3.1.4 TDF-QL

TDF and QL capture the relationships between the user query and a facet-value pair from different aspects and may complement each other. We combine these two features to score a facet-value pair as follows:

$$score(f:v,q) = \lambda * NS_l(P(q|f:v)) + (1-\lambda) * NS_l(tdf(f:v,q,N)) \tag{4}$$

where $NS_l(*)$ is to normalize the features. We use linear normalization here:

$$NS_l(s) = \frac{s - \min_{s_i \in \mathbf{S}}(s_i)}{\max_{s_i \in \mathbf{S}}(s_i) - \min_{s_i \in \mathbf{S}}(s_i)} \tag{5}$$

where $\mathbf{S}$ is the set of scores of all considered facet-value pairs.

3.2 Incorporate faceted feedback into retrieval

We present two retrieval models to incorporate user faceted feedback in this section. $\mathbf{P}_u$ denotes the set of facet-value pairs chosen by the user.

3.2.1 Boolean Model

The Boolean model filters documents with user faceted feedback. We can use the AND operation to require the retrieved documents contain all of the user-selected facet-value pairs. In practice, the AND operation might be too strict. One alternative is to use the OR operation to allow any document that contains at least one user-selected facet-value pair to pass. Another alternative is to use AND across different facets and OR within each facet. The Boolean model itself returns a document set instead of a ranked list. We can use any ranking methods, such as TFIDF, BM25 [21], etc., to rank the passed documents. We score documents by the Boolean model as follows:

$$s_{bl}(d) = \begin{cases} s_m(d) & \text{if } d \text{ contains all(AND) / any one(OR)} \\ & \text{facet-value pair } f:v \in \mathbf{P}_u \\ -\infty & \text{otherwise} \end{cases} \tag{6}$$

where $s_m(d)$ is the score of document d computed using a baseline ranking method m.

3.2.2 Soft Model

Despite the fact that the Boolean model is commonly used in the e-commerce domain, it may not work well for semi-structured text document retrieval. The Boolean model is based on two assumptions: 1) users are very clear about what they are looking for, and thus are able to select perfect facet-value pairs to restrict the returned documents; 2) document facets are accurate and complete so that no potentially relevant document is filtered out in retrieval due to meta data errors. These two assumptions may not hold in text document retrieval.

In a specific domain, some facets might be more informative than others. For example, for news articles, the information of time, locations, persons, and topics may be more important than publishers; for research papers, the subjects and keywords may be more informative than the file formats; For movies, the genres, casts and directors may be more informative than producers.

Based on the above motivations, we propose a soft retrieval model. In this model, we learn a weight for each type of facet, which is expected to reflect the quality of the facet. Here the quality may include user acquaintance, metadata accuracy, facet importance, etc. The soft model scores a document as follows:

$$\begin{aligned} s_{sm}(d) =& NS_s(s_m(d)) \\ &+ \sum_{f \in \mathbf{F}} \alpha_f * NS_l\Big(\sum_{f:v \in \mathbf{P}_u} \delta(d, f:v) * idf(f:v)\Big) \end{aligned} \tag{7}$$

where

$$\delta(d, f:v) = \begin{cases} 1 & \text{if } d \text{ contains } f:v \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

α_f is the weight of facet f and is learned automatically. $s_m(d)$ is the original score of document d.

$NS_s(*)$ is the standard normalization that converts the original document scores into a distribution with mean 0 and variance 1. The distributions of original document scores ($s_m(d)$) across different queries and using different baseline retrieval models might be significantly different. We found the difference would badly hurt the retrieval performance in our experiment. So we chose to normalize the original document scores first. As defined in Equation 5, $NS_l(*)$ is the linear normalization of the score part of a facet f.

4. EXPERIMENTAL METHODOLOGY

4.1 Datasets

To evaluate the proposed faceted feedback mechanism, we use two TREC filtering track datasets: the medical article collection OHSUMED and the news story collection RCV1 [14]. We choose these two corpora because they contain metadata, user queries/profiles, and relevance judgments.

OSHUMED dataset contains 348,566 medical articles selected from a subset of 270 medical journals covering years from 1987 to 1991. This dataset was used in the TREC 2000 filtering track [16], and we use the topics of this track to simulate user information needs in our experiment. The metadata field MeSH (Medical Subject Headline) is used as a document facet.

RCV1(Reuters Corpus Volume 1) dataset contains about 810,000 Reuters news stories published from 1996-08-20 to 1997-08-19. There are three types of codes assigned to documents in this collection: topic, geographical region, and industry. These codes are generated with a process involved a combination of auto-categorization, manual editing, and manual correction. We use the three codes as document facets. RCV1 was used in the TREC 2002 filtering track [17], and the first 50 topics of this track are used to simulate user information needs in our experiment[1].

4.2 Evaluation Based on Mechanical Turk

We use the Mechanical Turk [1] to collect user faceted feedback. Mechanical Turk is an online marketplace for work, where requesters can publish some tasks that require human intelligence, and workers can choose to work on the tasks to get paid. Comparing TREC assessors with Mechanical Turk workers, prior research shows Mechancial Turk workers are a good source for IR evaluation [3]. In our experiments, we ask workers to act as a real user to provide faceted feedback. For each query[2], we design a question, in which the TREC topic statement (including the title and description) and a group of recommended facet-value pairs are shown (See figure 1). Mechanical Turk workers are asked to select good facet-value pairs to restrict the search results according to their understanding of the query. The topic statement helps them act as if they are the real search engine users with the information need. By configuring the Human Intelligence Task (HIT) properties, we make sure there are three workers work on each query to give faceted feedback. Note that those workers are all random workers on Mechanical Turk who happen to see our task and choose to work on it. We also design some questions about their knowledge background related to the query topics in order to help us understand if there is a strong correlation between user knowledge levels and feedback quality.

[1]The prior research shows that the other topics do not match real user information needs well.

[2]Each query corresponds to a TREC topic.

4.3 Experimental settings

Our experiment is designed to answer the following questions:

- Is faceted feedback mechanism effective in improving retrieval performance?
- How does faceted feedback compare to other feedback mechanisms?
- Can faceted feedback be used together with other feedback mechanisms?
- Which facet-value pair recommendation methods are better?

To answer the first question, we compare the retrieval performance of faceted feedback to the baseline method BM25 without user feedback. In the baseline retrieval, only the title parts of TREC topics are used in order to simulate the short queries in real scenarios. To answer the second question, we compare faceted feedback with pseudo relevance feedback (PRF) and real document relevance feedback (RRF). We use the relevance judgments provided by TREC as simulated user feedback for RRF. To answer the third question, we use PRF and RRF respectively to calculate the original document scores ($s_m(d)$ in equation 6 and 7). The final retrieval performances will tell us whether faceted feedback complements existing feedback mechanisms and can be combined with them to further improve the retrieval performance. To answer the fourth question, we compare the retrieval performances of different faceted-value pair recommendation methods. Standard IR evaluation measures *Mean Average Precision* (MAP), *Precision@N* (P@N) and *Recall@N* (R@N) are used to evaluate the retrieval performances.

In our initial experiment, we found different users might choose different facet-value pairs given the same query and the same candidate set, which will lead to different retrieval performances. To avoid the influence of user difference, one possibility is to have the same user work on each of the candidate sets to be compared. However, a user's choice is influenced by his/her past experience and thus the order of how the candidate sets are presented will influence the results. To alleviate this problem, we combine the candidate sets recommended by four methods and present the large set to the user for feedback. When calculating the retrieval performance of a specific recommendation method, we only use the user-selected facet-value pairs that are included in the candidate set recommended by this method.

We set the number of facet-value pairs each recommendation method recommends (K) to 10, the number of top ranked documents used in the recommendation of facet-value pairs (N in equation 1) to 100, and the weight of query likelihood (λ in equation 4) to 0.5.

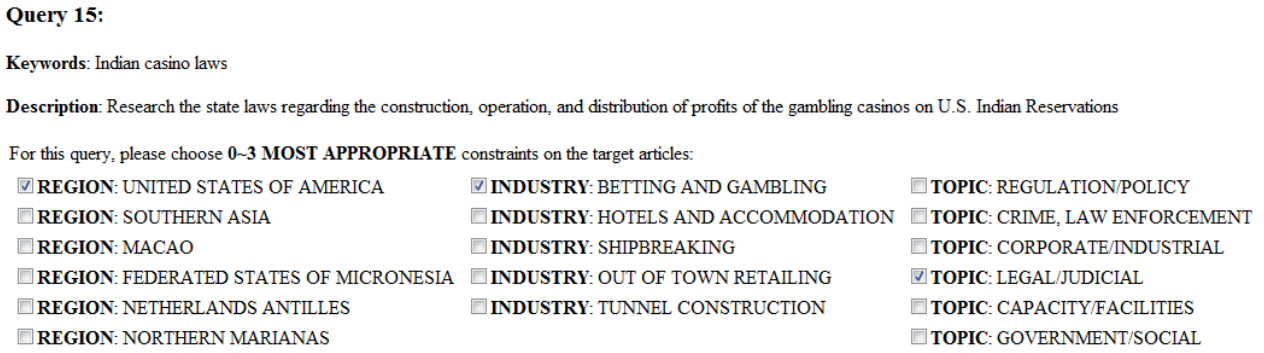

Figure 1: User interface on Mechanical Turk

5. EXPERIMENTAL RESULTS

5.1 Overall performances of faceted feedback

Table 1 shows the retrieval performances of the baseline (BM25), using faceted feedback (FF) from individual user (User1, 2 and 3 for OHSUMED dataset, User4, 5 and 6 for RCV1 dataset), and the average over three users (FF(Average)). P@10 is the precision of top 10 documents. All the performances reported here are obtained using the soft retrieval model[3]. The average MAP and P@10 of using faceted feedback on OHSUMED dataset are improved by 32.4% and 43.9% over the baseline (BM25) respectively. The average MAP and P@10 on RCV1 dataset are improved by 11.1% and 8.8% respectively. According to these results, we conclude that faceted feedback is effective in improving retrieval performance.

Table 1: Performances of Faceted Feedback (FF). "FF (User1|4)" means to use faceted feedback from User1 (on OHSUMED dataset) and User4 (on RCV1 dataset).

Dataset	OHSUMED		RCV1	
Performance	**MAP**	**P@10**	**MAP**	**P@10**
BM25 (baseline)	0.0921	0.1397	0.2907	0.5680
FF (User1|4)	**0.1354**	**0.2286**	0.3221	0.6180
FF (User2|5)	0.1112	0.1873	0.3150	0.6120
FF (User3|6)	0.1189	0.1873	**0.3318**	**0.6240**
FF (Average)	0.1219	0.2010	0.3230	0.6180
Imprv over BM25	32.4%	43.9%	11.1%	8.8%

5.2 User disagreement on faceted feedback

Given the same query and the same group of facet-value pair candidates, users may select different facet-value pairs, which lead to different retrieval performances. In Table 1, the performance using feedback from User1 and User6 (in bold) are better than other users. Table 2 gives two query examples for which users' faceted feedback are different from each other. Further analysis shows that there are very few queries that users gave exactly the same feedback. This is common in IR evaluation, as well trained TREC assessors usually have disagreements about document relevance judgments. This is also consistent with our anticipation: users' feedback may be different due to their different backgrounds and different understandings about the same information need. For example, users majoring in medicine are very likely to give more accurate feedback than the average user, which will result in better performance on the OHSUMED dataset. However, this does not mean faceted feedback is only useful for smart or expert users. Table 1 shows that three users' feedback are all useful in improving retrieval performances.

[3]The Boolean model will be discussed in a later section

Table 3: Performance comparison of different retrieval models on OHSUMED dataset. The feedback from User1 is used.

Retrieval model	MAP	P@10	R@1000
BM25 (baseline)	0.0921	0.1397	0.4612
Boolean model (AND)	0.0403	0.1522	0.0935
Boolean model (OR)	0.1120	0.1758	0.4650
Soft model	0.1354	0.2286	0.5301

Table 4: Performance comparison of different retrieval models on RCV1 dataset. The feedback from User6 is used. "Boolean model (A+O)" means to use AND operation across facets and OR operation within each facet.

Retrieval model	MAP	P@10	R@1000
BM25 (baseline)	0.2907	0.5680	0.6658
Boolean model (AND)	0.1046	0.3311	0.1514
Boolean model (A+O)	0.2208	0.5102	0.5062
Boolean model (OR)	0.2912	0.5780	0.6563
Soft model	0.3318	0.6240	0.6954

5.3 Boolean model v.s. Soft model

Table 3 and 4 compare the performances of the Boolean models and the soft model. R@1000 is the recall of top 1000 documents. In Table 4, "Boolean (A+O)" means to use AND across facets and OR within each facet (Table 3 doesn't have this since only one facet is used on the OHSUMED dataset).

Table 2: Examples of user-selected facet-value pairs

Query	User1	User2	User3
"58 yo with cancer and hypercalcemia"	MeSH:Hypercalcemia MeSH:Diphosphonates MeSH:Calcium	MeSH:Hypercalcemia MeSH:Carcinoma, Squamous Cell MeSH:Parathyroid Hormones	MeSH:Hypercalcemia MeSH:Paraneoplastic Syndromes MeSH:Bone Neoplasms MeSH:Bone Resorption
"Aborigine health"	Industry:Hospitals & Healthcare Topic:Health	Region:Australia Topic:Health Topic:Welfare, Social Services	Region:Australia Topic:Health Topic:Government/Social

The Boolean model with AND operation works poorly on both datasets. It results in much lower Recall@1000 than other retrieval models. The Boolean OR operation works better than the baseline method on OHSUMED dataset and a little worse on RCV1 dataset. The Boolean A+O works better than Boolean AND while still worse than Boolean OR. This reveals that when we loosen the Boolean restriction, we are actually getting improved retrieval performances. In contrast to the general practice of using Boolean approach in faceted search, the Boolean model in our experiments doesn't work well for text document retrieval. We did some further analysis and figured out two reasons for that. First, document metadata assignments are not perfect. Many documents are not assigned with metadata that they should have (we call this case incompleteness of metadata assignment). Secondly, some users select ambiguous or inappropriate facet-value pairs, probably because they are not familiar with the current topic. When using the Boolean model, many potentially relevant documents are filtered out due to either incompleteness of metadata assignment or users' inappropriate feedback, and thus the system recall is hurt seriously.

Soft model works well since it uses user feedback as preferences instead of rigid requirements. We proposed to use the parameter α_f to capture the quality of a facet previously. The motivation is that the values of some facets are easy to determine by either human beings or algorithms, while for some other facets this might be hard. For example, the facet "Region" might be easier for human, while "topic" might be harder. Someone may think a news article talking about "resident health" should be categorized into the topic "Government/Social" while some others may not think so. We found that the feedback on the "Topic" facet are different across three users, and feedback on the "Region" facet are more consistent across the users. Besides, for easier facets, the metadata assignment tends to be accurate and complete, and thus trustable. While for harder facets, the metadata assignment might be inaccurate and incomplete, and thus less trustable. These observations further justified our motivation for introducing parameter α_f.

The proposed soft model requires training data to learn α_f for each facet, and we use the 3-fold cross validation in our experiment. The queries are randomly split into three equal-size sets. In each fold, two sets are used as training queries to learn the parameter (α_f), and the last set is used for testing. The average performances over three folds are reported in Table 3 and 4. Table 5 shows the α values learnt in each fold[4]. On RCV1 dataset, α_{Region} are consistently larger than α_{Topic} and $\alpha_{Industry}$, which suggests that the "Region" facet is more trustable or easier for users than the other two facets.

Table 5: The optimal α_f trained in each fold

Dataset	Fold	Facet	Optimal α
OHSUMED	Fold1	MeSH	8.5
	Fold2	MeSH	7.5
	Fold3	MeSH	9.5
RCV1	Fold1	Region	10
		Topic	2
		Industry	1.5
	Fold2	Region	10
		Topic	1.5
		Industry	0.5
	Fold3	Region	2.5
		Topic	2
		Industry	2

It is worth mentioning that if a software such as those proposed by Herst et al.[11] is used to automatically generate facet values, we may prefer completeness/recall of metadata assignment instead of precision. Because metadata incompleteness hurts the system recall badly, while inappropriate facet value assignment hurts less, as the final ranking algorithm would rank a document low if it is non-relevant to the user information need.

5.4 Comparison of different facet-value pair recommendation approaches

Figure 2 compares the retrieval performances corresponding to four facet-value pair recommendation methods: TDF (Top N Document Frequency), TDFIDF (Combine TDF with IDF), QL (Query Likelihood), and TDFQL (Combine QL with TDF). The horizontal axis shows the number of facet-value pair candidates used, and the vertical axis shows the corresponding MAP. The performances shown in the figures are the average across three users. TDFIDF and TDFQL methods perform better than the other two methods on OHSUMED dataset. This is consistent with our expectation, since both methods combine two features of facet-value pairs.

Interestingly, on RCV1 dataset, TDFIDF and TDFQL perform worse than TDF. Further analysis shows that the

[4]The smallest scale we tried for α_f is 0.5, since smaller scales have no significant influences on retrieval performances.

facet-value pair "Region:USA" benefits retrieval performance a lot for several queries. Unfortunately, both TDFIDF and TDFQL rank it out of the top 10 since it appears frequently in the whole corpus, so users have no chance to see this candidate. One possible solution is to use facet weights (α_f) as an extra feature in our facet-value pair recommendation methods. The motivation is that those facets more trustable for retrieval (with bigger α_f, such as the facet "Region" over the other two facets) should be boosted in recommendation. This is consistent with the second criterion of good facet-value pairs we mentioned in section 3.1. We will evaluate this idea in our future work.

5.5 Comparison with other types of feedback

The retrieval performances of faceted feedback (FF), Pseudo Relevance Feedback (PRF) and Real document-based Relevance Feedback (RRF) are compared and shown in Table 6. The performances of combining PRF with FF (PRF@5+FF) and combing RRF with FF(RRF@5+FF) are also reported. The performance of FF is the average over three users. The top 5 ranked documents in BM25 are used for pseudo relevance feedback (PRF@5). The relevance judgments of top 5 are used in the BM25 relevance feedback algorithm as implemented in Lemur [2] (RRF@5). Table 6 shows that FF performs better than PRF, and closely to RRF on OHSUMED dataset; FF performs worse than PRF and RRF on RCV1 dataset, and 10% better than BM25.

Though FF might perform worse than RRF, FF is still very promising because of three major reasons. First, a retrieval system may get more faceted feedback than document feedback, as faceted search is commonly accepted by the average Internet user and faceted feedback seems very easy for a user to understand. Second, RRF and PRF often help little for hard queries when no relevant documents are retrieved in the top positions, while faceted feedback might help boost relevant documents in these cases. A major scenario where search engine fails is that an engine only focuses on one aspect of a query and ignores some other important aspects [7]. Faceted feedback provides the mechanism for users to put constraints on the important aspects to avoid this problem. Actually, in our experiment, we find there are a number of queries for which FF helps a lot when PRF and RRF help little or even hurt. These queries are mostly hard queries with poor initial retrieval performances. Take the query "Nuclear plants U.S." as an example, almost all the initially returned top documents in baseline ranking are about nuclear plants outside U.S., thus PRF and RRF hurt. FF helps since all users are able to identify the faceted restriction "Region:USA", which boosts those documents talking about events happening in U.S.. Third, different types of feedback are not exclusive and they could complement each other. We can easily combine FF with PRF or RRF and obtain better retrieval performances. This can be done by using PRF or RRF as the baseline method to calculate the original document scores ($s_m(d)$ in equation 6 and 7). Table 6 shows that the combininations of FF with PRF or RRF improve the performances further.

Table 6: Performance comparison of different types of feedback. FF: faceted feedback; PRF@5: pseudo relevance feedback using top 5 docs; RRF@5: real document-based relevance feedback using top 5 docs.

Dataset	OHSUMED		RCV1	
Performance	**MAP**	**P@10**	**MAP**	**P@10**
BM25 (baseline)	0.0921	0.1397	0.2907	0.5680
FF	0.1219	0.2010	0.3230	0.6180
PRF@5	0.1096	0.1746	0.3711	0.6280
RRF@5	0.1240	0.2048	0.3887	0.6940
PRF@5+FF	0.1269	0.1937	0.3899	0.6320
RRF@5+FF	0.1473	0.2481	0.4025	0.7007

6. CONCLUSIONS

We researched the user feedback mechanism based on faceted document metadata. The results on a medical dataset and a news dataset show that faceted feedback is useful, though different users may give different feedback for the same query. Directly using the Boolean model, which is commonly used in e-commerce, is inappropriate for metadata-based general purpose document retrieval, since the document metadata assignment is usually incomplete. The proposed soft model is shown consistently more effective on both datasets, as it automatically learns a weight for each facet, which captures the facet quality. The proposed facet-value pair recommendation methods are generally effective and can be improved in the future. Faceted feedback could be combined with pseudo relevance feedback and document relevance feedback. We tried one simple combining method and found better retrieval performance.

In the future, more research is needed to explore different facet-value pair recommendation algorithms, for example, incorporating facet weights (α_f), considering the interaction among facet-value pairs and how user choices are affected by context. We also want to explore different ways to combine various feedback mechanisms.

7. ACKNOWLEDGMENTS

We thank Jessica Gronski and Yize Li for valuable discussions related to this research. The work was funded by National Science Foundation IIS-0713111, AFRL/AFOSR and UCSC/LANL Institute for Scalable Scientific Data Management. Any opinions, findings, conclusions or recommendations expressed in this paper are the authors', and do not necessarily reflect those of the sponsors.

8. REFERENCES

[1] Amazon mechanical turk. https://www.mturk.com.

[2] The lemur toolkit for language modeling and information retrieval. http://www.lemurproject.org/.

[3] O. Alonso and S. Mizzaro. Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, 2009.

[4] P. Anick. Using terminological feedback for web search refinement: a log-based study. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 88–95, New York, NY, USA, 2003. ACM.

[5] P. Anick and S. Tipirneni. Interactive document retrieval using faceted terminological feedback. *Hawaii*

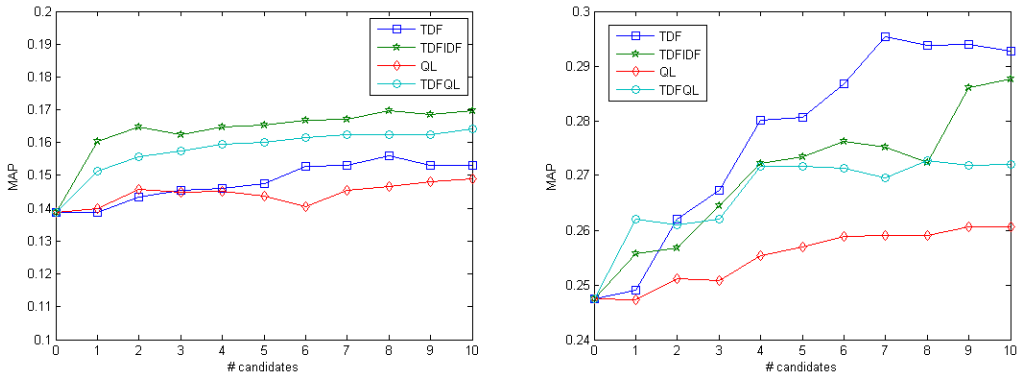

Figure 2: Performances of different facet-value pair recommendation approaches. The left figure: on OHSUMED dataset; the right figure: on RCV1 dataset

International Conference on System Sciences, 2:2036, 1999.

[6] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *SIGIR*, 1998.

[7] C. Buckley. Why current ir engines fail. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 584–585, New York, NY, USA, 2004. ACM.

[8] B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information Retrieval in Practise*, chapter 11 Beyond Bag of Words, page 463. Pearson, 2009.

[9] J. C. French, A. L. Powell, F. Gey, and N. Perelman. Exploiting a controlled vocabulary to improve collection selection and retrieval effectiveness. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 199–206, New York, NY, USA, 2001. ACM.

[10] D. Harman. Towards interactive query expansion. In *SIGIR '88: Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–331, New York, NY, USA, 1988. ACM.

[11] M. A. Hearst and E. Stoica. Nlp support for faceted navigation in scholarly collection. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 62–70, Suntec City, Singapore, August 2009. Association for Computational Linguistics.

[12] D. Kelly, V. D. Dollu, and X. Fu. The loquacious user: a document-independent source of terms for query expansion. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 457–464, New York, NY, USA, 2005. ACM.

[13] D. Kelly and X. Fu. Elicitation of term relevance feedback: an investigation of term source and context. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 453–460, New York, NY, USA, 2006. ACM.

[14] D. D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. 2004.

[15] H. Raghavan, O. Madani, and R. Jones. Active learning with feedback on features and instances. *J. Mach. Learn. Res.*, 7:1655–1686, 2006.

[16] S. Robertson and D. Hull. The TREC-9 filtering track report. In *The Ninth Text REtrieval Conference (TREC-9)*, pages 25–40. National Institute of Standards and Technology, special publication 500-249, 2001.

[17] S. Robertson and I. Soboroff. The TREC-10 filtering track final report. In *Proceeding of the Tenth Text REtrieval Conference (TREC-10)*, pages 26–37. National Institute of Standards and Technology, special publication 500-250, 2002.

[18] J. J. Rocchio. Relevance feedback in information retrieval. 1971.

[19] X. Shen and C. Zhai. Active feedback in ad hoc information retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 59–66, New York, NY, USA, 2005. ACM.

[20] L. Si and R. Jin. Flexible mixture model for collaborative filtering. In *ICML '03: Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[21] S. J. M. H.-B. Stephen E. Robertson, Steve Walker and M. Gatford. Okapi at trec-3. 1994.

[22] B. Tan, A. Velivelli, H. Fang, and C. Zhai. Term feedback for information retrieval with language models. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 263–270, New York, NY, USA, 2007. ACM.

[23] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. pages 403–410, 2001.

[24] Y. Zhang. Using bayesian priors to combine classifiers for adaptive filtering. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 345–352, New York, NY, USA, 2004. ACM.