

A User-Centered Approach to Evaluating Topic Models

Diane Kelly¹, Fernando Diaz², Nicholas J. Belkin³, and James Allan²

¹SILS, University of North Carolina
Chapel Hill, NC, USA 27599
dianek@ils.unc.edu

²CIIR, University of Massachusetts, Amherst
Amherst, MA, USA 01003
[fdiaz, allan]@cs.umass.edu

³SCILS, Rutgers University
New Brunswick, NJ, USA 08901
nick@belkin.rutgers.edu

Abstract. This paper evaluates the automatic creation of personal topic models using two language model-based clustering techniques. The results of these methods are compared with user-defined topic classes of web pages from personal web browsing histories from a 5-week period. The histories and topics were gathered during a naturalistic case study of the online information search and use behavior of two users. This paper further investigates the effectiveness of using display time and retention behaviors as implicit evidence for weighting documents during topic model creation. Results show that agglomerative techniques - specifically, average-link clustering - provide the most effective methodology for building topic models while ignoring topic evidence and implicit evidence.

1 Introduction

A general problem for current interactive information retrieval (IR) systems is disambiguating the topic of interest to a searcher, given a statement of the person's information problem, typically posed as a rather brief query. One possible approach to this issue is to take advantage of the person's previous information seeking behaviors in order to identify topics which have been of interest to that person in the past. This could be done, for instance, by recording the documents (e.g. Web pages) that the person has looked at as a result of searching for information, and automatically classifying those pages according to topic models, derived from the language of the documents. A new search by the person could be associated with one or a few of such models, thereby effectively disambiguating the search topic, and providing a basis for searching for new documents which might be generated by the topic model(s). Language modeling and clustering techniques have proven useful for generating topic models in other domains [1,18]. However, the effectiveness of such techniques on personal collections has yet to be tested.

Another method of topic identification is to observe such behaviors as display (or dwell) time, or bookmarking, printing or otherwise saving or using documents. Given such evidence from previous and current behaviors, documents of current interest could be related to documents of past interest, and therefore to topic models.

The purpose of this paper is to explore a novel method of evaluating the accuracy of topic models which have been created using traditional language modeling and clustering approaches, and behavioral evidence, such as display time and retention. This method consists of comparing topic models created using these approaches with those created by users during a naturalistic study using self-identified topics.

2 Related Literature

Recently, language modeling has received much attention in the IR community [5]. In this framework, a collection of text data—documents or sets of documents—is considered to be sampled from an underlying generative process. Namely, we assume that the words in a document were all generated according to some topic model. A topic model is a probability distribution over words. For example, the topic “Iraq War” may have some probability of generating the terms “Bush”, “Hussein”, and “Iraq” and much lower probability of generating “cattle”, “dance”, and “salsa”.

Figure 1 provides a graphical interpretation of this approach. A topic language model can be described using the urn metaphor. Each topic is considered an infinite collection of words which follow some distribution. A document is produced by iteratively picking words from this urn. Unfortunately, we do not have access to the true distribution of terms in this topic urn, but known, on-topic documents can be used as evidence for estimating this distribution. Fortunately, casting the task as model estimation allows one to use formalisms from statistics. Previous techniques such as the vector space model often relied a great deal on heuristics and hand-tuned parameters.

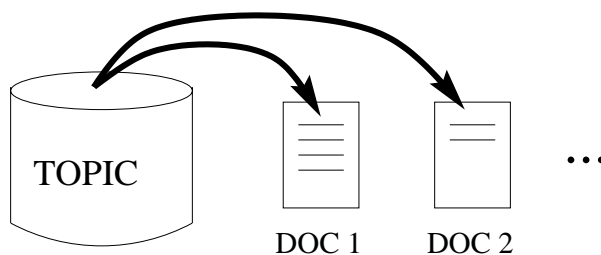


Fig. 1. Topic Language Model: A topic can be interpreted as an infinitely large collection of words. Documents are generated by choosing words from this urn.

The most relevant work for topic modeling has been conducted in the context of Topic Detection and Tracking [1]. This literature deals with the problem of tracking several topics in a stream of news articles. The community has produced several

techniques for automatically clustering and detecting links between documents in a stream. Although more sophisticated language modeling systems have been developed for these tasks, the most successful approaches use straight-forward vector-space techniques [2]. Nevertheless, language modeling systems provide a formal methodology for estimating the topic models.

Several techniques have also attempted to create user-topic clusters in an IR setting. In these cases, clusters are constructed by incorporating explicit user feedback usually in an interactive IR setting [3,6,8]. In an environment of passive feedback, many authors have described the incorporation of disambiguating terminology from a user's search history [7,15]. These systems often build a user model and leverage this information to expand the query.

Other approaches to user modeling for personalized IR have used the user's online behaviors as implicit indicators of interest [4,10,12,14,16]. Typical approaches have used display time, retention (e.g. printing, saving, and bookmarking), scrolling and selection to identify relevant documents for feedback during a single search session. Behavioral evidence has also been used to cluster search results [9], but it has not been used as evidence for topic model construction.

The evaluation of topic models has typically consisted of various cluster or link detection measures [1]. These approaches often use annotator consensus as a baseline for evaluation. It is often difficult to assess how a technique will actually perform in "real-life" because of a lack of a user-centered evaluation metric. Instead, assumptions must be made about how a user would classify, label and evaluate documents.

In the paper, we evaluate the accuracy of two language model-based clustering techniques with user-defined topic classes of web pages from personal web browsing histories. We further investigate the effectiveness of including behavioral evidence in the construction of these models.

3 Monitoring Study

The data used for the study reported in this paper was collected during a naturalistic case study of the online information-seeking behaviors of two users during a five-week period. Users were provided with laptop computers and their activities were monitored with logging and evaluation software and online questionnaires.

We chose a naturalistic approach because we were interested in providing users with an opportunity to engage in multiple information seeking episodes over time, with tasks and topics that were germane to their personal interests, in familiar searching environments. The naturalistic approach also provided an advantage over web server log analysis because the identity of users could be maintained and various measurements could be collected during the observational period. Furthermore, information about intentions and specific tasks and topics could be gathered and associated with behaviors and documents. We chose to conduct two descriptive case studies because we were interested in gathering a large, detailed quantity of data. We do not claim our two users to be a sample, nor do we claim that our results generalize reliably to a larger population of users.

3.1 Users

Two volunteer users completed the five-week study. Both users were graduate students in a Master's of Library and Information Science program and held Bachelor's and Master's degrees in the Humanities. Both users had a high degree of self-assessed computer and online searching experience.

3.2 Instruments and Procedures

Each user was provided with an IBM ThinkPad equipped with the Windows 2000 operating system and standard utilities to use for the duration of the study. The laptops were equipped with client-side logging software that monitored and recorded users' interactions with the operating system and all other applications. The monitoring software was launched automatically each time the laptop was started, executed in stealth mode while the laptop was in operation and recorded information such as applications used, URLs visited, start, finish and elapsed times for interactions and all keystrokes. A proxy server captured all pages the user viewed while connected to the Internet.

Two types of behaviors were of interest to this study: display time and retention. In this study, display time was the length of time that a document was displayed in the user's active browser window. Display time was collected from the client-side logger, which indicated elapsed times for displaying a particular document. Since the client-side logger recorded active window data and all programs with which the user was interacting, we feel somewhat confident that this measure was accurate. While we cannot insure that the user was viewing the document and not attending to other off-line activities, we are confident that display time data collected from a client-side logger is more reliable than that collected from a proxy server. Retention behaviors [13] included saving, printing, emailing or bookmarking, and were gathered directly from the client-side logger.

At the beginning of the study, users read and signed a consent form, which outlined the protocol of the study and informed users that all of their activities with the laptops would be monitored. An Entry Questionnaire, which elicited background information from the user, such as education and search experience, was administered. A Task and Topic Questionnaire was also administered that elicited the tasks and topics the user would be engaged with during the period of the study. Users were asked to think about their online activities in terms of tasks and topics. For example, a task might be shopping and the topic of this task might be clothing, or guitars. Another example task might be writing a research paper; the topic of this task might be political ecology and West Africa.

All pages that the user viewed while searching were captured by a proxy server. A content-based classification of page types was created based on a manual examination of 2000 pages to identify and select systematically the pages that were to be evaluated in the study. The goal was to eliminate pages such as ads, search pages, email pages, etc. Two independent coders validated this classification.

A Task and Topic Update Questionnaire was administered online each week of the study, which presented users with their previously identified tasks and topics and asked them to update the list through additions and/or deletions.

At the mid- and end-points of the study, the pages viewed up to that time were presented online to the users for evaluation. The instrument used for this evaluation displayed the text of one page at a time, a console which had two drop-down lists containing the user's tasks and topics and text boxes in the event that new tasks or topics needed to be added during the evaluation. Users were asked to classify each page that they viewed according to its task and topic and to evaluate the usefulness of the page using seven-point usefulness (1=not useful, 7=useful) and confidence scales (1=low, 7=high).

3.3 Results of the Monitoring Study

There were several types of data that we were interested in using from the monitoring study. We were interested in the users' self-identified topics and their classification of the documents that they viewed into each of these groups. We were interested in users' display time and retention behaviors. Finally, we were interested in the usefulness ratings that users associated with each page. In sum, the data from the monitoring study provided us with sets of documents that had been clustered into self-identified topics by users, and usefulness scores and behaviors for each document. In this study, we did not consider the task classes created by our users. While this information may be helpful in distinguishing topic classes, we leave it for future analysis.

A total of 2353 items were logged by the proxy for User 1 and 533 were logged for User 2. After screening the documents according to the classification described above, 427 (18%) were identified for evaluation by User 1 and 198 (36%) for User 2. Table 1 displays an overall description of the number of documents viewed and evaluated by each user, the number of topics identified, the mean usefulness and confidence of the pages evaluated and the mean display times. Interestingly, the mean display time for all documents for both users was identical. In general, both users were very confident with their evaluations of the documents that they viewed.

Table 1. Description of behavior and page evaluations

	User 1	User 2
Documents Viewed	2353	533
Documents Evaluated	427	198
Topics Identified	20	15
Usefulness (Mean, SD)	5.0 (1.03)	4.28 (1.96)
Confidence (Mean, SD)	6.03 (.33)	6.21 (.78)
Display Time (Mean, SD)	0:53 (2:33)	0:53 (2:24)

Users identified a range of topics with which they were engaged throughout the study. A list of topics for each user is displayed in Table 2. This table includes the number of documents viewed for each topic (\bar{D}), the mean display time of documents

for each topic (RT), the number of retention behaviors for each topic (RET) and the mean usefulness (Use). Many of these topics were related to libraries, since both users were Masters students in library and information science, and both users worked in libraries. Also, both users were concurrently enrolled in the same course, and several topics were related to the same course project. For example, both users identified the topic of “evaluation criteria,” which was related to a course assignment about developing evaluation criteria to assess the usability of web resources. “Review material” and “book review research” also represented a particular course project that required students to identify and evaluate online sources for book reviews. Other topics represented users’ specific content-based interests in libraries; User 1 studied theology-based texts, while User 2 studied classic texts and various materials associated with classic texts such as papyri. Other topics represented individual interests unrelated to the course, such as “eyeglasses,” “sailing,” and “recipe search.”

We used all user-defined topic clusters in the present study as a baseline with which to evaluate clusters created using automatic techniques.

Table 2. Topics identified by User 1 and User 2 and characteristics of each

User 1					User 2				
Topic	D	RT	RET	Use	Topic	D	RT	RET	Use
Theology	55	00:38	1	4.45	General Interest	57	01:12	4	5.52
Perennials	11	01:00	0	5.00	College Financial Aid	5	00:23	0	3.40
North Carolina	2	00:23	0	5.00	Papyrology, palaeography, epigraphy	14	00:36	1	4.92
Library Literature	116	01:20	21	4.65	New York City	1	00:47	0	6.00
Review Material	33	00:32	4	4.94	Classics	24	00:39	0	3.54
Homestead Rebate	1	00:27	1	5.00	Woodcarving	18	00:29	0	2.35
Eyeglasses	7	00:38	1	5.14	Mass Transit	17	00:30	0	3.12
Weddings	53	01:12	1	4.65	Directions	2	00:19	0	3.50
Rescued Beagles	101	00:27	3	5.85	Serials	5	00:56	0	5.20
Poison Ivy	6	02:25	0	4.83	Medieval	6	00:32	0	5.67
Evaluation Criteria	5	00:58	0	4.40	Electronic Resources	1	02:38	0	7.00
Florida	3	00:25	0	5.00	Collection Development	2	00:16	0	5.00
U. of Arizona	4	00:53	0	5.00	Recipe Search	6	05:36	0	4.67
Alexander Library	3	00:15	0	5.67	Evaluation Criteria	3	00:28	1	1.67
Classmate	5	00:16	0	6.20	Book Review Research	37	00:25	1	3.62
Amanda Beasley	5	00:43	0	5.00					
Sailing	9	00:55	0	5.22					
Dog Park	4	00:23	1	4.00					
Radio	3	00:52	0	5.33					
Music	1	00:10	0	4.00					

4 Topic Clustering Techniques

This section contains a description of the various techniques used to cluster the documents viewed by our users, including the rules we used to identify useful documents. This section is followed by a description of the techniques that were used to compare the clusters of documents generated by our users with the clusters generated by the statistical approaches.

We take a language modeling approach to modeling collections of text [5]. Assume we are given a single document as a sample from the urn described in Figure 1. A naïve estimation of the document model would merely count the frequencies of the terms in the document,

$$\hat{P}(w | D_i) = \frac{c(w, D_i)}{\sum_{w \in D_i} c(w', D_i)} \quad (1)$$

where $c(w, D_i)$ represents the number of times word w appears in a particular document, D_i . Unfortunately, if a word does not occur in the document, then its estimated probability will be zero. Since a document is only a sample from this document model, we would like to smooth this estimate with some other model. We accomplish this by interpolating the maximum likelihood document model with a maximum likelihood collection model so that our smoothed document model becomes,

$$P(w | D_i) = \lambda \hat{P}(w | D_i) + (1 - \lambda) \hat{P}(w | C) \quad (2)$$

where C is the document collection. In our experiments, λ was empirically set to 0.90. *Topic* language models can then be constructed by combining the individual document language models. For instance, if we know that a set of documents T all discuss the same topic, then we build the topic language models according to,

$$P(w | T) = \frac{1}{|T|} \sum_{D_i \in T} P(w | D_i) \quad (3)$$

This formalism can be used to build concise summaries of topics by inspecting the language model for each topic. Specifically, we can compute how the topic model, $P(w | T)$, differs from the collection model, $\hat{P}(w | C)$, by inspecting the *pointwise Kullback-Leibler (KL) divergence* [7,17]. For each word, the pointwise KL divergence is defined as,

$$P(w | T) \log \frac{P(w | T)}{\hat{P}(w | C)} \quad (4)$$

Terms with the highest pointwise KL divergence will be the most discriminating.

These topic models are estimated by using *all* of the evaluated documents and serve not as a method to be evaluated but rather to qualitatively represent the language modeling technique. For example, Table 3 displays the top ten distinguishing words

for the topics identified by our users. Notice that it is not necessarily the case that topics with fewer example documents have less meaningful language models. It is more important to have a consistent and precise language. An example of where consistency and precision fail for large topic sizes can be seen in the language model description of User 2's topic "General Interests".

Table 3. Language model (LM) descriptions of user topics

User 1		User 2	
Topic	LM Description	Topic	LM Description
Theology	Biblical, bible, Israel, theology	General Interest	Weather, jesus, movie, film, time, home
Perennials	Geranium, plant, garden, big	College Financial Aid	Loan, pay, hesc, forbearance, borrow
North Carolina	Weather, forecast, low, high	Papyrology, palaeography, epigraphy	Citation, library, abstract, full, article
Library Literature	Library, information, service	New York City	Square, greenmarket, union, park
Review Material	Title, library, book, footage	Classics	Classic, rate, site, resource, ancient
Homestead Rebate	Taxation, treasury, rebate, state	Woodcarving	Nantucket, art, carve, stbart, gallery
Eyeglasses	Store, lenscrafter, offer, rate	Mass Transit	Transit, rail, corridor, transportation
Weddings	Wedding, bride, indiebride, club	Directions	Switchboard, starbuck, map, search
Rescued Beagles	Pet, petfinder, beagle, dog, org	Serials	Edit, reprint, und, von, teil, die, der
Poison Ivy	Hive, webmd, cause, post	Medieval	Der, kehr, papsturkunden, paul, fridolin
Evaluation Criteria	Evaluate, internet, site, web	Electronic Resources	Rom, text, edition, database, English
Florida	Mapquest, map, flight	Collection Development	Record, franco, view, gesta, dei, nogent
U. of Arizona	Semester, session, class, summer	Recipe Search	Recipe, chicken, epicurious, cook, sauce
Alexander Library	Rutgers, library, alex, summer	Evaluation Criteria	Hon, honcode, medical, health
Classmate	Yahoo, map, locate, glen, address	Book Review Research	Book, review, booklist, june
Amanda Beasley	Elliot, ilisha, nerve, feature		
Sailing	Race, Bermuda, Newport		
Dog Park	Maplewood, construct		
Radio	Wfuv, wny, folk, stream, city		
Music	Garbage, 22garbage, band, google		

We are interested in automatically recognizing and representing topics in the pages viewed and evaluated by our users. Two clustering methods were implemented for automatically building topic models: k-means clustering and agglomerative clustering.

4.1 K-means Clustering

In our context, k-means clustering assumes that there are k underlying topics responsible for having generated the documents in the training data. The learning process begins by randomly picking k documents as cluster representatives or *centroids*. The remaining documents are then assigned to the most similar topic model. For our experiments, similarity is determined by the Kullback-Leibler divergence defined as

$$KL(D_i || T) = \sum_{w \in V} P(w | D_i) \log \frac{P(w | D_i)}{P(w | T)} \quad (5)$$

where V is the vocabulary. After all documents have been assigned, topic models are then re-estimated using the new document topic sets. This assignment and estimation process continues until topic models converge. For each user, k was set to the known number of topics.

Selection of Seed Documents. We assume that in real interaction there will be documents already associated with particular topics. Therefore, we next consider how to feed examples to this algorithm. Our approach is to assign these examples to the initial centroids and fix these assignments throughout the execution of the learning. Therefore, the example documents will always be a component of the topic model.

The method of selection for these seed documents consisted of identifying documents receiving the highest usefulness rating (7) and the highest confidence rating (7) by our users for each topic. In cases where more than one document met the selection criteria, an attempt was made to select documents that were viewed on different days. In cases where only a single day was represented, the first two documents meeting the criteria were selected. If there were no documents for a specific topic class with a usefulness rating of 7, then documents that received a 6, the second highest usefulness rating, were selected. In most cases the confidence scores were always high, so it was possible to select documents which had high confidence scores associated with them.

Display Time and Retention. In addition to providing seeds for the topic models, we also considered weighting documents depending on their import. In particular, we were interested in using the display time and retention behaviors of our users as implicit evidence of usefulness. The goal in doing this was to unobtrusively identify documents whose weight could be increased during the automatic clustering process.

For many topics, there were only a few documents viewed. Because of this, we used a measure of display time based on the overall display times and usefulness ratings for each user, rather than those display times observed for individual topics. We grouped the points of our 7-point usefulness scale into 3 classes: Low (1-3), Medium (4) and High (5-7). We then computed the mean display time for each of these groups for each user and used the mean display time for the high usefulness group as a method for identifying useful documents. Thus, if a user displayed a document for longer than this mean display time, then the weight of this document was increased during clustering. The means for each usefulness group are displayed in Table 4. Our use of retention was a little more straightforward. If a retention behavior occurred at a

document (i.e. the user printed, saved, emailed or bookmarked the document), then we used this to increase the weight associated with the document during clustering.

Table 4. Mean and standard deviation display times according to usefulness

	Usefulness Group		
	Low (1-3)	Medium (4)	High (5-7)
User 1	00:28 (00:23)	00:48 (03:12)	00:57 (02:35)
User 2	00:21 (00:21)	00:35 (01:03)	01:22 (03:23)

4.2 Agglomerative Clustering

One drawback to the k-means approach is the requirement that we know the number of topics, k . As an alternative, we also evaluated two agglomerative clustering techniques. Agglomerative clustering techniques build topic representations bottom up. The algorithm begins with each document in its own cluster and then successively merges clusters according to similarity. The method always merges the two closest clusters. It is the interpretation of *closest* which differentiates our two agglomerative techniques. In both cases, clustering terminates when the similarity between the closest clusters is below a certain threshold.

Single-link Clustering. One possible interpretation of inter-cluster distance considers the shortest distance between all inter-cluster document pairs (i.e. a document belongs to the same topic as its most similar neighbor). For this algorithm to be consistent, we use the J-divergence, a symmetric version of the KL-divergence measure,

$$J(D_i \parallel D_j) = KL(D_i \parallel D_j) + KL(D_j \parallel D_i) \quad (6)$$

It is important to notice that the single-link technique provides no explicit representation of a topic model. Because of this, a method for seeding the algorithm with topic examples is not obvious and was not used for this technique.

Average-link Clustering. Although single link clustering performs well in traditional topic tasks, it has a tendency to create topic models covering a variety of sub-topics; this is a product of a document only needing a single highly similar match to be included in the cluster. Instead, we may want to assign a document to the cluster to which it has the highest *average* similarity. In this case, the similarity between two clusters is calculated by averaging the similarity between all pairs of documents between two clusters.

5 Evaluation Techniques

We used as ground truth the clusters that resulted from our users' classification of the documents that they viewed into self-identified topics. The automatically-generated clusters were evaluated by measuring the accuracy of predicted links between docu-

ments. That is, two documents in the same cluster are said to have a link between them. If there are N documents for a particular user, then there are $O(N^2)$ possible links between all pairs of documents. Let this total set be L . Let the set of true links defined by the manual clustering of the documents be defined by $L' \subseteq L$. Let the set of links predicted by the system be defined by $L_s \subseteq L$. We evaluate the performance of our systems using two measures of accuracy. First, we measure the *total accuracy* of prediction,

$$\frac{|L' \cap L_s|}{|L'|} + \frac{|\overline{L'} \cap \overline{L_s}|}{|\overline{L'}|} \quad (7)$$

This evaluates the system prediction of link presence and absence in a set of documents. Our second measure focuses on the accuracy of predicting true links. Specifically, we use the equation,

$$\frac{|L' \cap L_s|}{|L'|} \quad (8)$$

which will provide a means for disambiguating the degree to which good *total accuracy* relies upon keeping unrelated documents in separate clusters. We will refer to this as *link recall*.

6 Results

Eight variants of the k-means clustering technique were used which incorporated different degrees of evidence and re-weighting. The agglomerative techniques were run without any evidence or re-weighting. Thresholds for the clustering were empirically set. Table 5 presents the total link accuracy and link recall for each subject, for each technique. The results presented in Table 5 indicate that seeding clusters provides valuable information for the k-means techniques. In all cases, seeded clusters out-perform the unseeded counterparts. However, other results for the k-means techniques are less conclusive. For example, the effect of re-weighting schemes such as display time on performance is mixed. We speculate that a more sophisticated incorporation in the k-means model might provide better results. Surprisingly, the knowledge-poor agglomerative techniques performed as well or better in three out of the four trials.

While these results provide gross estimates of system performance, we would like to measure the actual number of true links retrieved. Table 5 displays the accuracy of predicting topical links between documents (link recall). Again, seeding and re-weighting improve performance. Further, the agglomerative techniques perform better than k-means for User 1. The agglomerative results for User 2 are less conclusive perhaps as a result of the small collection size.

We speculate that the poor performance of the k-means experiments for User 1 is the result of fixing k . If the language of the documents does not follow a topical pattern, then restricting the potential cluster assignments will result in conflating

distantly related documents. The agglomerative techniques are quite content leaving those outliers as singleton clusters, effectively remaining agnostic about topic assignment. This is confirmed by the large number of singleton clusters in the agglomerative techniques. These statistics are shown in Table 6.

Table 5. Total link accuracy and link recall

	User 1		User 2	
	Accuracy	Recall	Accuracy	Recall
K-means, no seeds				
No re-weighting	0.600633	0.318978	0.553381	0.200471
Display time	0.59475	0.293309	0.562427	0.189707
Retention	0.609477	0.361913	0.553381	0.200471
Both	0.564198	0.201446	0.567767	0.212244
K-means, seeded				
No re-weighting	0.692427	0.426086	0.59018	0.242852
Display time	0.690125	0.4212	0.591452	0.24588
Retention	0.694866	0.428236	0.59018	0.242852
both	0.693766	0.427976	0.591452	0.24588
Single Link	0.717979	0.47723	0.572933	.260343
Average Link	0.723693	0.471301	0.612068	.214262

Table 6. Number of clusters generated by agglomerative techniques

	User 1	User 2
Single Link	151	33
Average Link	85	32

In order to further test this hypothesis, additional k-means experiments were performed with alternate values for k . No seeded experiments were performed because there would be fewer seeds than clusters. After sweeping a range of k from 21 to 50, a value of 30, in general, improved the unseeded performance the most compared to the original experiments. The results for these experiments are shown in Table 7. Note that only the un-weighted and display time-weighted techniques actually improved with an increase in k . In fact, the performance of methods incorporating retention is, in general, worse when we increase the number of clusters.

Table 7. 30-means experiments for User 1, no seeds

	Total accuracy	Link recall
No re-weighting	0.664936	0.385823
Display time	0.664987	0.385758
Retention	0.563404	0.200078
Both	0.564102	0.198059

These results indicate that the agglomerative techniques are more successful for User 1 not because of their superior representation but rather because they take fewer risks in deciding that two documents are on the same topic. The k-means algorithms, on the other hand, are potentially forced to take these risks. The benefit is that if the language encodes the topics, accuracy of known links is better than the agglomerative techniques.

7 Discussion

For both users, the average-link technique provided the most accurate performance. The k-means technique without seeds performed the worst, even when display time and retention were considered. In all cases, seeded clusters outperformed the unseeded counterparts and models of User 1's topics were more accurate than User 2's. Including behavioral evidence for re-weighting documents resulted in little, if any, improvement. When considering link recall, the agglomerative techniques outperformed the k-mean techniques and the seeded clusters still outperformed the unseeded.

The use of display time and retention had little effect on clustering with or without seed documents. It may be the case that our measure of identifying useful documents based on display times was not the most effective. Previous work [11] has found that while documents that are rated more highly usually have higher mean display times, they also have higher variance, which might make it difficult for measures of mean display times to perform very well. An examination of Table 4 demonstrates that this may be the case for this data as well. Thus, our measure of usefulness based on display time may not have been selective enough. Moreover, User 1 rated a large portion of the documents that she viewed as 5 or 6, which are points included in the high usefulness group. While our use of mean display time is an improvement over previous work since its computation is based on the behavior of each individual user as opposed to a group of users, it still may not have been sensitive enough.

Retention was found to increase performance slightly for User 1 when no seeds were used (Table 5), perhaps indicating its potential as a technique for identifying documents that could be used as seeds or for re-weighting. User 1 exhibited more retention behaviors (33 documents) than User 2 (7 documents), which may explain why clustering for User 1 benefited from the inclusion of retention for re-weighting.

Set size and quality may have also affected our results. In terms of total number of documents evaluated, User 1 viewed and evaluated considerably more than User 2, while only identifying 5 more topics. It should be noted again that our users did not evaluate all of the documents that they viewed during the 5-week period. Instead, we screened documents using a classification scheme to eliminate email pages, advertisements, discussion groups and search pages. Thus, we believe that the quality of the documents that were evaluated and used in the clustering, were better than if we had used all displayed documents. Although we cannot be certain without conducting the analysis, clustering only the set of documents that users evaluated most likely resulted in more accurate topic models than clustering all displayed documents.

The number of documents users associated with each topic varied considerably. For some topics, 50 or more documents were associated with the topic, but it was more often the case that a large number of topics had 5 or fewer documents associated with them. Given that we used 2 seed documents per topic, it is unsurprising that the k-means with seeds out-performed no seeds for many topics.

8 Conclusions

Overall, the techniques we used for topic model construction performed poorly when evaluated according to the user-defined topic classes. It is unclear if automatic clustering techniques can be as sensitive as users when creating and assigning documents to topic clusters. However, we feel that more attempts at user-centered approaches to the evaluation of topic models are necessary and that the clusters created by users can provide an evaluation metric of the highest standard. Moreover, great care was invested in developing the methodology used in our monitoring study, and we believe that this methodology can act as a valuable model for others interested in exploring user-centered approaches to evaluating automatically-generated topic models.

A combination of the quality of the sets and the seeds seems to play a significant role during clustering. The use of seeds improved performance for both users and suggests that the identification of quality seeds may be necessary for accurate topic modeling. Additionally, the quality of the documents that were evaluated and used in the clustering were better than if we had used all displayed documents. Certainly the definition of a “quality” document is rather nebulous and more work needs to be done understanding and identifying the attributes of quality documents. If we are to create a system that makes use of a user’s web browsing history, then the system needs to know when it should consider a document for inclusion in topic clustering. Whether relevant or not relevant, not all documents are equally useful in constructing topic models. A document that only contains a search box is not as useful as one which contains the text of a conference paper. This also applies to using behavior as implicit feedback: observing a high display time at a document containing a search box and little text most likely indicates something different than observing a high display time at document containing a conference paper. Clearly, the system needs some assistance in identifying candidate documents for inclusion in topic modeling and as sources of implicit feedback. We are currently working to develop our web page classification for use in future experiments and hope that this will elucidate some aspects of “quality” documents and their impact on modeling.

We have just finished a second naturalistic study of the sort described in this paper with seven new users, which lasted 3.5 months. We plan to conduct an analysis and evaluation similar to the one described in this paper and adjust our display time measure, as well as investigate the usefulness of the task groupings and additional behavioral data in the construction of topic models. Ultimately, we would like to use these models to provide personalized information retrieval to individuals.

9 Acknowledgments

This work was partially supported by the CIIR, NSF Grant #IIS-9907018, and NSF Grant #99-11942. Opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsor. We would like to thank Doug Riecken and IBM for the loan of the laptops and Haizheng Zhang for the proxy logger.

References

1. Allan, J., (Ed.) Topic Detection and Tracking, Event-based Information Organization, Kluwer Academic Publishers, 2002.
2. Allan, J., Lavrenko, V., & Swan, R. Explorations within topic tracking and detection. In Topic Detection and Tracking, Event-based Information Organization (James Allan, Ed.), Kluwer Academic Publishers, 197-224, 2002.
3. Bhatia, S. K. & Deogun, J. S. Cluster Characterization in Information Retrieval. Proceedings of SIGIR '93, 721-728, 1993.
4. Claypool, M., Le, P., Waseda, M., & Brown, D. Implicit interest indicators. Proceedings of Intelligent User Interfaces (IUI '02), 2001.
5. Croft, W.B. & Lafferty, J., (Eds.) Language Modeling for Information Retrieval, Kluwer Academic Publishers, 2003.
6. Deogun, J.S. & Raghavan, V.V. User-oriented Document Clustering: A framework for learning in information retrieval. Proceedings of the ACM SIGIR '86, 157-163, 1986.
7. Diaz, F. & Allan, J., Browsing-based User Language Models for Information Retrieval, CIIR Technical Report IR-279, 2003.
8. Gordon, M. User-based document clustering by redescribing subject descriptions with a genetic algorithm. JASIST, 42, 311-322, 1991.
9. Heer, J., & Chi, E. H. Separating the swarm: Categorization methods for user sessions on the web. Proceedings of CHI '02, 243-250, 2002.
10. Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L.R. & Riedl, J. GroupLens: Applying collaborative filtering to usenet news. Communications of the ACM, 40(7), 77-87, 1997.
11. Kelly, D. & Belkin, N. J. Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback during interactive information retrieval. Proceedings SIGIR '01, 408-409, 2001.
12. Morita, M., & Shinoda, Y. Information filtering based on user behavior analysis and best match text retrieval. Proceedings of SIGIR '94, 272-281, 1994.
13. Oard, D. W., & Kim, J. Modeling information content using observable behavior. Proceedings of ASIST '01, 38-45, 2001.
14. Pazzani, M. & Billsus, D. Learning and revisiting user profiles: The identification of interesting web sites. Machine Learning 27, 313-331, 1997.
15. Ruthven, I., Lalmas, M., van Rijsbergen, K. Incorporating user search behavior into relevance feedback. JASIST, 54, 529-549, 2003.
16. Seo, Y. W., & Zhang, B. T. Learning user's preferences by analyzing web-browsing behaviors. Proceedings of Autonomous Agents, 381-387, 2000.
17. Tomokiyo, T. & Hurst, M. A language model approach to keyphrase extraction. Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition & Treatment, 33-40, 2003.
18. Zhai, C., Cohen, W.W., & Lafferty, J. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. Proceedings of SIGIR '03, 10-17, 2003.

Kelly, D., Diaz, F., Belkin, N. J., & Allan, J. (2004). A user-centered approach to evaluating topic models. In *Proceedings of the 26th European Conference on Information Retrieval (ECIR '04)*, Sunderland, UK, 27-41.