# Evaluating Distance-Based Clustering for User (Browse and Click) Sessions in a Domain-Specific Collection

Jeremy Steinhauer[1], Lois M.L. Delcambre[1], Marianne Lykke[2], Marit Kristine Ådland[3]

[1]*Dept. of Computer Science, Portland State University, Portland, OR, U.S.A*

{jsteinha, lmd}@cs.pdx.edu

[2]*Dept. of Communication and Psychology, Aalborg University, Aalborg, Denmark*

mlykke@hum.aau.dk

[3]*Dept. of Library and Information Science, Oslo University College, Oslo, Norway*

marit-kristine.adland@hioa.no

**Abstract.** We seek to improve information retrieval in a domain-specific collection by clustering user sessions from a click log and then classifying later user sessions in real-time. As a preliminary step, we explore the main assumption of this approach: whether user sessions in such a site are related to the question that they are answering. Since a large class of machine learning algorithms use a distance measure at the core, we evaluate the suitability of common machine learning distance measures to distinguish sessions of users searching for the answer to same or different questions. We found that two distance measures work very well for our task and three others do not. As a further step, we then investigate how effective the distance measures are when used in clustering. For our dataset, we conducted a user study where we had multiple users answer the same set of questions. This data, grouped by question, was used as our gold standard for evaluating the clusters produced by the clustering algorithms. We found that the observed difference between the two classes of distance measures affected the quality of the clusterings, as expected. We also found that one of the two distance measures that worked well to differentiate sessions, worked significantly better than the other when clustering. Finally, we discuss why some distance metrics performed better than others in the two parts of our work.

## 1 Introduction

With the advent of the Internet, collections often allow searching and browsing. And sites often have logs that capture browse moves in addition to queries and click-throughs. We are interested in using sessions in such a log from domain-specific sites to make recommendations or personalize search results to improve information retrieval.

Researchers have attempted to make recommendations or personalize search results based on profiles, ratings, and web usage logs. Many researchers have used machine learning to cluster users based on the similarity of their behavior [1-10]. Based on the cluster to which a user belongs, some predict items in a collection that might be of interest to that user (collaborative filtering). However, we have found no studies that directly address the fundamental viability of using click logs with these machine learning techniques. In particular, few have evaluated the overall performance of the clustering, much less reported the effects of various distance measures or ways of representing user actions. In this paper we address the suitability of machine learning techniques to cluster web usage logs via these research questions:

1. How well can we distinguish between users searching for answers to the same questions from users searching for answers to different questions using distance measures?
2. Does question similarity affect the ability to tell these two types of sessions apart?
3. Does question difficulty (measured by session length) affect the ability to tell them apart?
4. When used in a clustering algorithm, does the ability of a distance measure to differentiate between sessions translate to better performance?
5. Which distance measure performs best for distance-based clustering of user session data?

Essentially we are investigating a variant of the cluster hypothesis proposed by Jardine and van Rijsbergen [11] which states, "the association between documents convey information about the relevance of documents to requests." Their work was comparing the similarity between documents based on the text of relevant documents to a query. We hypothesis that the association between users' sessions (the documents viewed) convey information about other documents that could be relevant to those users.

To answers these questions we need sessions from multiple users answering the same, similar, and different questions. We conducted a user study where we asked participants to find the answer to a set of questions of varying degrees of difficulty and similarity and recorded the pages they clicked on while attempting to answer the questions.

In the first phase of our work, we used distance measures to compute the distance between pairs of sessions answering the same, similar and different questions using four different session vector representations. This evaluation is similar to the evaluation Voorhees used in testing the cluster hypothesis [12] and allowed us to answer our first three research questions. Significantly, we found two classes of distance measures: ones that could discriminate between sessions answering the same vs. similar or different questions (namely, cosine vector and Tanimoto) and ones that could not (namely, Euclidean, squared-Euclidean, and Manhattan). We found that the amount of similarity between questions affected our ability to differentiate questions but that we could still observe a significant difference between sessions of users answering the same vs. similar questions. Question difficulty (average session length)

for the poorly performing class of distance measures was surprisingly positively correlated with average distance between sessions; though, this correlation seems to be more of a property of the distance measure than the data. While this data allowed us to see the discriminating power of distance measures, it did not allow for a direct comparison of distance measure. This is because the various distance measures produce distances in various ranges and scales.

Since we intend to use distance-based clustering algorithms to cluster users' sessions for later classification, as the second phase of our work, we chose to investigate whether the observed differences in distance measures would affect distance-based clustering. We compared the resulting clusters using extrinsic cluster quality metrics; this allowed us to directly compare the performance of distance measures. To compute the extrinsic cluster quality metrics, we compared clusterings from the clustering algorithms to our gold standard: sessions clustered based on the question they were answering.

As a preliminary step for phase two, we considered clustering algorithms of four different types: k-Means (a centroid-based algorithm), single link hierarchical (a linkage-based, bottom-up, hierarchical algorithm), OPTICS (a density based algorithm), and EM (a probabilistic based algorithm with no distance measure involved). We found that for our data, k-Means provided the best overall results.

As part of phase two, we investigated what effect varying numbers of clusters had on performance and found that the optimum number of clusters is between 80 and 120 for our study. Overall, we confirmed that there were two classes of distance measure, but in the top performing group, Tanimoto distinguished itself as the best for our data and task.

This paper extends an earlier paper that reported the results from phase one of this work [11].

This paper is organized as follows. We describe related work in Section 2, the methods we used for gathering and analyzing our data in Section 3, and our data analysis in Section 4. In Section 5, we offer conclusions and describe what can be done to build on our work.

# 2 Related Work

Much of the work in applying clustering algorithms in the information retrieval field has focus on clustering documents based on the content of the pages to find similar pages. Early papers in this field [11, 12] describe the problem and propose basic tests for evaluating the potential of these techniques for improving the state of the art. While our work is focused on the potential of machine learning techniques using web usages logs, many of the tests proposed in this earlier work are applicable.Strehl et al. [13] systematically analyzed distance measures for the purposes of grouping similar documents. They clustered documents based on the content of the pages, using a similar set of distance measures and clustering algorithm as our work. They found similar results regarding distance measures used in clustering: cosine and extended Jaccard (Tanimoto) similarity are vastly superior to other measures. The main difference be-

tween our research and theirs is that they focused on clustering based on page content not user sessions. Also, their work investigated distance measures after clustering, whereas we also compared distance measures to a gold standard directly.

Some researchers have considered applying clustering algorithms to web usage logs. But, much of this work fails to evaluate the ability of their algorithms to associate users with similar information needs. Also many researchers focus purely on searches and documents clicked in search results (click-throughs) but do not consider a user's full search and browse history [1-8,15].

Some work has used web usage logs to analyze users' query/click-through and browse behavior. Ageev et al. [16] gathered logs of users answering a set of questions that were deemed hard to answer. They were interested in classifying sessions to identify successful searching behaviors. Their study did not make use of clustering algorithms or distance measures and their logs involved web wide searches whereas we limit ourselves to a single web site.

Mobasher et al. [17] compared the performance of several clustering algorithms using web usage data for the purpose of personalized page recommendation. They clustered a portion of their log and then classified sessions from the rest of their log to their clusters, using a portion of the pages from each session for classification and the rest as a relevance judgment. Similar to us they found that k-means performed the best. Our work differs from theirs in several ways: we have a gold standard, we know what each user was searching for during their session; and we investigate the role distance measures, session vector representation, questions similarity, and question difficulty play in clustering user sessions.

# 3 Methods

The methods used for the first phase of our work are described in Sections 3.1 through 3.4. In Sections 3.1 and 3.2 we describe how we selected and analyzed questions and how we used those questions in our user study. We describe the session vector representations we used in the first phase of our study in Section 3.3 and the distance measures we used for both phases in Section 3.4. The clustering algorithms and extrinsic cluster metrics we used in the second phase of our study are described in Section 3.5.

## 3.1 Questions

We selected the American Cancer Society's website (cancer.org) as our domain-specific collection. We gathered 141 questions from cancer forums, the question and answer sections of cancer sites, and question-asking sites such as Yahoo! Answers. We determined that 120 of these questions could be answered using cancer.org by manually finding the answers on the American Cancer Society site.

We wanted to use questions with a range of similarities and difficulties in our study. A lay person estimated the difficulty of each question, on a scale from 1 to 5, based on how long it took to find the answer using cancer.org. To determine related-

ness, we had an oncologist list which cancer types, if any, were associated with each question. We also had a lay person identify the general topics that appeared in the questions; 15 topics were chosen. Zero or more topics were associated with each question, as appropriate. For example the question: "Where do ampullary cancers normally start?" was associated with the topic *detection*.

To quantify the distance between questions for the purposes of choosing questions and performing our analysis, we used two question distance measurements. The first question distance measure, $QDist_{tfidf}$, compares term vectors for the text of a question concatenated with the associated cancer types (from the oncologist) and the associated topics (from the lay person). Stop words were removed and the Porter stemming algorithm was applied. Each position in the vector represented a term; there was a position for each unique term found in the questions, cancer types and topics. A value > 0 in a position in the vector indicated that the question had that term associated with it. Term vectors were weighted using the TF-IDF score for each term. The cosine distance measure was used to determine the distance between the term vectors for each question pair. The $QDist_{tfidf}$ score is on a scale from 0 to 1 where 0 is the same and 1 is completely different.

The second question distance measure was based on the percentage of overlap of related cancer types between a pair of questions. If $T_1$ and $T_2$ are the cancer types associated with question $Q_1$ and $Q_2$ and $C=T_1 \cap T_2$, the question distance score $QDist_{ct}$ is:

$$QDist_{ct}(Q_1,Q_2) = 1 - |C| / ((|T_1| + |T_2|) / 2). \qquad (1)$$

For this measure, 0 means that the questions have the same associated cancer type(s) and 1 means they have no cancer types in common.

In order to prepare a list of questions with a mix of similar and different questions for the users who participated in our study, we used our first method for measuring question distance. (The second method for measuring distance between questions was used only during analysis of our results.) Question pairs with a score from 0 to .65 exclusive were considered similar and from .85 to 1 were considered different. These boundaries were set such that the algorithm could pick 40 questions while maximizing the gap between the scores of the two groups. Our highest scoring (i.e., least similar) pair that was still deemed similar with distance score of .63 was:

- What is retinoblastoma?
- In what age range is retinoblastoma most commonly found?

Our lowest scoring pair deemed different, with a similarity score of .86, was:

- Can one still have children after testicular cancer?
- Can chemotherapy or radiation cause anemia?


**Fig. 1.** Depiction of how questions were placed into groups. Color represents groups of similar questions with black representing questions with no similarity.

We programmatically selected 40 questions (4 groups of 10) to use in our user study. To test a range of similarities, we selected three sets of four similar questions, three sets of three similar questions and three sets of two similar questions (27 in total), such that within a set the questions were similar but were not similar to any of the other questions in the set of 27. Then we chose 13 questions that were not similar to any other question. We used a greedy algorithm that chose questions from the larger sets of similar questions, then the smaller sets, and finally the set of different questions. Questions were put into 4 groups of 10, as show in Fig. 1, such that within each group all questions were pairwise dissimilar to prevent a training effect from users knowing where to find information. Each group had at least one question from each of the 5 difficulty levels.

### 3.2 User Study

Each participant in our study was given one of our groups of 10 questions to answer using only cancer.org and only the interface we provided. We used a proxy server to present cancer.org and captured the user's click stream.

We used Amazon's Mechanical Turk to recruit 200 participants. Google's CAPTCHA test was used to ensure that the participants were people. The participants were randomly divided into 4 groups of 50. Each group was given one of the predefined groups of 10 questions. Participants were presented questions one at a time, in random order. They also were given a frame set to the homepage of cancer.org at the start of each new question to be used for answering the question. Participants were given 45 minutes to answer the questions. They were paid one dollar (an amount on the high end for Mechanical Turk compensation [18]) for following the rules and participating. The top 25% of users, based on the number of correct answers given, earned an additional dollar. We based our study on one by Ageev et al. [16] which was designed as a question answering game to encourage participation; the incentive payment creates competition which encourages effort and makes it a game. Correct answers were determined ahead of time by searching cancer.org. We expanded our definition of what constituted a correct answer, as appropriate, by examining the final page of users' sessions to see if we could find a correct answer. (We were fairly lenient; correctness of an answer was only used for determining compensation).

We eliminated sessions for which there was only one page hit other than the homepage (the first page in every session) unless the correct answer was obviously on that page. Such users may have already known the answer, used an outside source, or simply guessed. We kept all other sessions even though users may have answered the question incorrectly or not at all.

We placed user clicks into individual sessions where one session consisted of the pages viewed by one user answering one question. We eliminated page clicks associated with the game, such as question submissions, and cancer.org's homepage since it appeared in every session. We standardized the escaping of characters in URLs and we analyzed URLs in our logs to determine pages that had different URLs but referred to the same page either by redirects or by standard URL conventions.

### 3.3 User session model: session vector representations

We modeled each session S as a vector of pages P where each position, $P_i$, in the vector represents one page in the set of all unique pages viewed in the course of the study (our corpus). We set the values at each position of the vector four ways:

*Binary* represents whether or not a page was viewed in the session.

$$\forall P_i \in P: Binary(P_i) = \begin{cases} 1 \; if \; the \; page \; was \; viewed \\ 0 \; otherwise \end{cases} \quad (2)$$

*Frequency* represents the number of times a page was viewed in the session.

$$\forall P_i \in P: Frequency(P_i) = \#\; of \; time \; P_i \; was \; viewed \; in \; S \quad (3)$$

*PFISF*, Page Frequency times Inverse Sessions Frequency, a weighting formula we defined based on TF-IDF, that gave a larger score to pages that have been viewed by fewer sessions. It takes the frequency value for $P_i$ from (3) and divides it by the number of sessions in which $P_i$ has appeared, $F_i$.

$$\forall P_i \in P: PFISF(P_i) = Frequency(P_i) * \frac{1}{F_i} \quad (4)$$

*Tail weighting* reflects the idea that later pages tend to be more important. We used a linear formula: the closer to the end of the user's sessions, the higher the weight. Let pos($P_i$,S) be the position of the page $P_i$ in the user's session S (if a page appeared more than once, the later page position is used).

$$\forall P_i \in P: Tail(P_i) = \; pos(P_i, S)/|S| \quad (5)$$

### 3.4 Distance Measurements

We investigated the ability to discriminate between pairs of session vectors, x and y, using the following standard machine learning distance measures, provided by Mahout, an open source machine learning tool from Apache:

- Cosine Vector (cos)

$$cos(x,y) = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n}(x_i{}^2)\sum_{i=1}^{n}(y_i{}^2)}} \quad (6)$$

- Euclidean (euc)

$$euc(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (7)$$

- Manhattan (man)

$$man(x,y) = \sum_{i=1}^{n}|x_i - y_i| \quad (8)$$

- Squared Euclidean (sqe)

$$sqe(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2 \quad (9)$$

- Tanimoto (tan)

$$tan(x,y) = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\left(\sum_{i=1}^{n}(x_i{}^2) + \sum_{i=1}^{n}(y_i{}^2)\right) - \sum_{i=1}^{n} x_i y_i} \quad (10)$$

These distance measures take two vectors and compute a distance score with 0 for identical vectors with increasing scores as the vectors become farther apart. Note, these distance metrics are not necessarily directly comparable; they provide distances in different ranges and with different distribution. In the first phase of our analysis, we evaluated the suitability of a distance measure and a session vector representation for discriminating between sessions answering the same, similar, or different questions by looking at the average distance, with standard deviation, among all pairs of questions answering the same question, among all pairs of questions answering similar questions, and among all pairs of questions answering different questions. We were particularly interested in whether the average distance among these three types of question pairs were different and whether their standard deviation provided a clean separation between the three types. Each distance measure was evaluated separately. This analysis is quite similar to that Voorhees used when testing the cluster hypothesis.

### 3.5 Clustering algorithms and Metrics for Clustering Quality

We conducted a preliminary test to choose which clustering algorithm(s) to use for our analysis of distance measures. We compared four clustering algorithms each representative of a different class of algorithm: k-Means (a centroid-based algorithm), single link hierarchical (a linkage based bottom-up hierarchical algorithm), OPTICS (a density based algorithm), and EM (a probabilistic algorithm with no distance measure involved). We used open source machine learning libraries from Mahout [19], Weka [20], and Elki [21] for implementation of these algorithms. We ran each algorithm five times using the cosine vector distance measure (except for EM which does not use a distance measure) and had each produce 40 clusters. We chose to use cosine vector because it was one of the top-performing distance metrics from phase one of our work and because it is widely used and commonly available in the machine

learning libraries. We chose to produce 40 clusters because our gold standard has 40 clusters corresponding to the 40 questions that users answered in our study. For OPTICS a specific number of clusters could not be set directly so we adjusted parameters such that we achieved the best performance for ~40 clusters.

For phase two of our study we investigated the effect distance measures had on the performance of a clustering algorithm. To evaluate and compare clusterings, we used extrinsic metrics based on a comparison of the clusters resulting from a clustering algorithm to our gold standard. Our gold standard consists of forty clusters corresponding to the forty original questions; each session was placed into the cluster corresponding to the question that was being answered. The metrics we used are based on true positive (TP), true negative (TN), false positive (FP), and false negative (FN) statistics produced by the comparison with the gold standard. For a pair of sessions, those answering the same question and in the same cluster are TPs, same question in different clusters are FNs, different questions in the same cluster are FPs, and different questions in different clusters are TNs.

We used the following metrics to evaluate a clustering:

- Precision (P)

$$\frac{TP}{TP+FP} \quad (11)$$

- Recall (R)

$$\frac{TP}{TP+FN} \quad (12)$$

- F Measure (with precision and recall equally weighted)

$$\frac{2PR}{P+R} \quad (13)$$

- Fowlkes-Mallows Index

$$\sqrt{PR} \quad (14)$$

- Rand

$$\frac{TP+TN}{TP+FP+TN+FN} \quad (15)$$

To test distance measures, we clustered using the best performing algorithm from our initial step, k-means, with the five distance measures we analyzed in the first phase of our study. We also tested the effect of varying the numbers of clusters, k, on performance by using k of 40, 80, 120, 160, and 200. We ran each permutation of distance measure and k five times, randomizing session order prior to each run. The first k sessions were used as the initial k points in the algorithm.

**Table 1.** Averaged data for questions used in the Mechanical Turk study

| | Avg. | Std. Dev. | Min | Max |
|---|---|---|---|---|
| **Sessions** | 50 | 0 | 50 | 50 |
| **Accepted** | 30.75 | 4.25 | 15 | 39 |
| **Time spent** | 68.2 s | 36.7 s | 7.6 s | 171.7 s |
| **Clicks** | 4.45 | 1.37 | 1.69 | 8.41 |

# 4 Results

Here we report basic statistics of the Mechanical Turk data set. Then we analyze the data with respect to our research questions.

## 4.1 Mechanical Turk Study Data

200 participants were recruited and completed our study in less than 4 hours. Table 1 shows averages per question from the Mechanical Turk study. We had a 61.50% acceptance rate (30.75/50) based on our criteria for an acceptable session, as described in Section 3. This acceptance rate is a little below what others have reported [16], but not atypical for Mechanical Turk studies.

Session lengths (clicks) varied with a minimum of 1.69 and a maximum of 8.41. We used session length to indicate question difficulty in our analysis; this data shows a range of difficulties as we had hoped.

## 4.2 Same vs. different questions

Our first research question considered whether standard machine learning distance measures could differentiate sessions of people answering the same questions from sessions of people answering different questions. We looked at our four session vector representations: binary, frequency, PFISF and Tail. Fig. 2 shows average pairwise distances grouped into same, similar, or different (based on $QDist_{tfidf}$ score for question distance) for each session vector representation and distance measure. Each bar represents the average distance between the pairs in a group. The color of the bar represents whether the pairwise distances between sessions were for sessions answering questions that are all: the same (same); similar, (0-.65] $QDist_{tfidf}$ score (sim); not the same (sim & dif); or different [.85-1] $QDist_{tfidf}$ score (dif). Session distances were normalized before averaging by dividing by the max value. Error bars in the figure represent one standard deviation. (Note, we refer back to this figure when we discuss the effect similarity has on our ability to differentiate similar questions in the next subsection.)

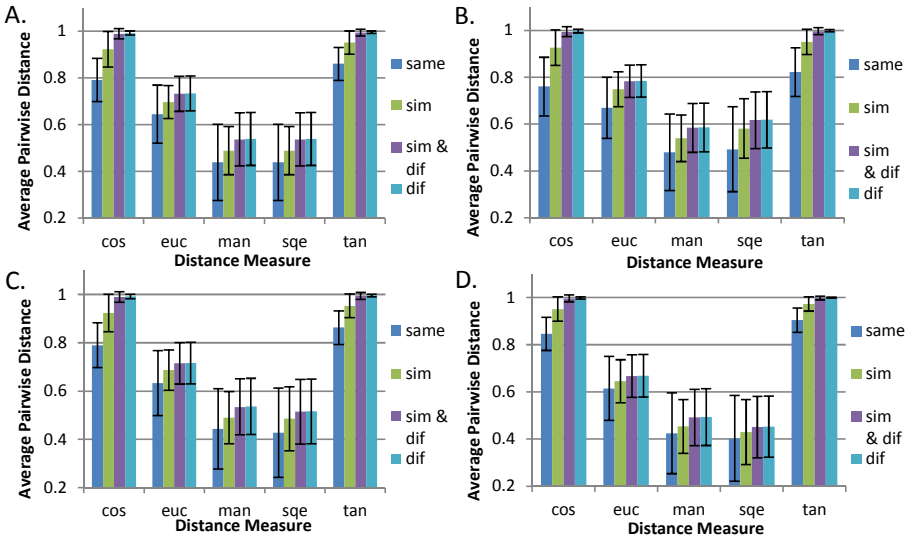


**Fig. 2.** Average pairwise session vector distances with standard deviation for each distance measure (cos, euc, man, sqe, and tan). The color of bars within each group represent whether sessions pairs were answering same (same), similar (sim), or different questions based on $QDist_{tfidf}$. Sim & dif is the union of similar and different session pairs. Different graphs represent different session vector representations: A. binary, B. tail, C. frequency, D. PFISF

To compare the distance between pairs of sessions answering the same question (same) and the distance between pairs of sessions not answering the same question (sim & dif) consider the first and third bar in each grouping. The distance measurements fall into two classes, regardless of session vector representation. For cosine vector (cos) and Tanimoto (tan), we see a clean separation between same vs. sim & dif questions; there is no overlap in the error bars. A T-Test comparing these two sets shows the difference is highly significant; with average scores of 2.29E-13. The combination of cosine vector distance and a frequency weighting perform the best with a score of 6.18E-16. The second class of session vector distance measures consists of Euclidean (euc), Manhattan (man), and squared Euclidean (sqe). For this group, we see nearly completely overlapping error bars with an average T-Test score of 0.0146. While this result is still significant it is 11 orders of magnitude less significant than the other group.

### 4.3 Similarity

Our second research question was: what effect does the similarity between questions have on our ability to differentiate between sessions from users answering the same question and sessions of users answering different questions?

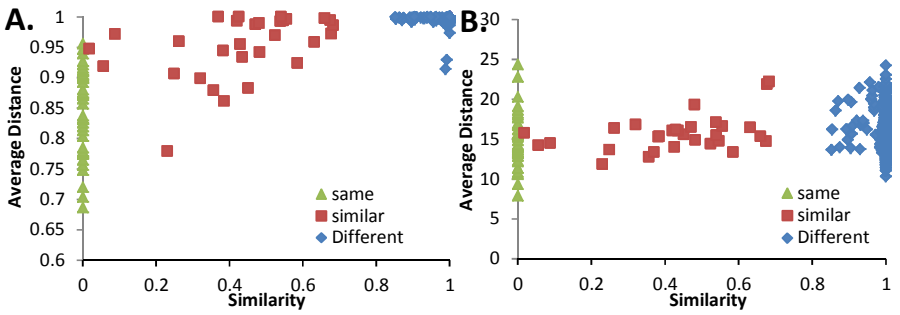


Fig. 3. Question distances, $QDist_{tfidf}$, compared with average session distances, A. cosine vector. B. Euclidean, each using the PFISF session vector representation.

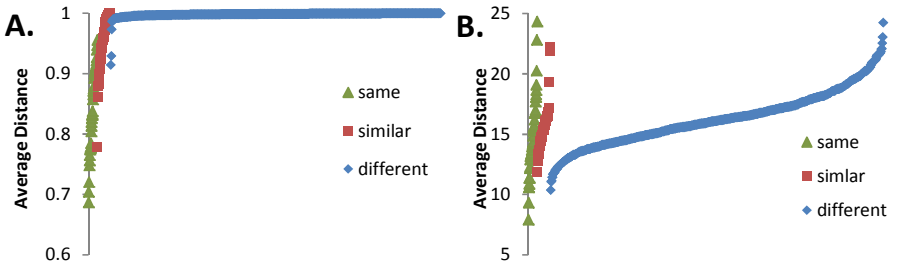


Fig. 4. Average session distances, A. cosine vector and B. Euclidean, each using the PFISF session vector representation and separated into $QDist_{tfidf}$ similarity groups and ordered by average session distance.

Look again at Fig. 2: in each grouping compare the second column (sim), the average distance between sessions for people answering similar questions, (0-.65] using $QDist_{tfidf}$ and the fourth column (dif), the average distance between sessions for people answering different questions, [.85-1] using $QDist_{tfidf}$. As one might expect, the average session vector distance between similar questions (sim) is closer in value to the average session vector distance between same questions (column 1), regardless of weighting or distance measure, than the average session vector distance between different questions(dif) is to the average session vector distance between same questions. We again see the same two groups of differently performing distance measures. For tan and cos, the error bars overlap; but the difference between same questions and similar questions is still strongly significant with an average T-Test score of 8.68E-5: Cosine vector and PFISF performed the best with 1.57E-5. For the other group (euc, sqe, and man), except for one instance, the error bar for sim is contained in the error bar for same, and there was no significant difference with T-Test scores averaging .309; the one outlier was euc with tail weighting which was barely significant, 0.0469.

Fig. 3 shows the average pairwise session vector distance between sessions for all pairs of questions plotted against their $QDist_{tfidf}$ scores: 0 for same, (0-.65] for similar, and [.85-1] for different. For the rest of the paper, we use cosine vector and Euclidean as representatives of the two classes of distance measures and we limit ourselves to the PFISF – the best performing session vector representation. In graph 3A (cosine vector), for similar question pairs, we see an upward trend as we move from left to

right. A line fitted to these points had a positive slope and an $R^2$ value of .191. While not significant, this result still suggests that as question pairs are deemed less similar, their sessions tend to be farther apart (have fewer page clicks in common).

Fig. 3A has some outliers: a red square and two blue diamonds that are well below the others. The question pair for the similar question outlier (red square) is:

- What is retinoblastoma?
- In what age range is retinoblastoma most commonly found?

These questions are about an uncommon cancer for which there is relatively little information on the cancer.org website. So it is not surprising that questions related to this cancer might hit the same pages.

The two outliers in the different question pair range (blue diamonds) reflect a shortfall of our TF-IDF distance metric. One question was in each of these two question pairs and the other two questions were considered similar (the common question and the other two were classified as different). The question in both pairs was:

- How long does it take for a normal cell to become cancerous after it starts changing?

The other two questions in the pairs deemed similar to each other were:

- When does a tumor become cancerous?
- Are all tumors cancerous?

We can see that all of these questions are quite similar, especially the question found in both pairs and the first of the other two. However, the only term that they have in common that is not a stop word is cancerous. This term stems to cancer which appears quite often in our questions and therefore has a low TF-IDF score. These are also general questions that do not have a specific cancer type associated with them; thus neither of our question similarity measures served to indicate that these questions are similar.

Fig. 3B clearly shows how the Euclidean distance measure had trouble differentiating between same, similar and different questions. All three groups of questions pairs (same, similar, different) have a similar range of distance scores. Fig. 4 shows the same data as Fig. 3 (average pairwise session vector distance of sessions for all question pairs) sorted from closest to farthest divided into three groups: same, similar and different indicated by color and shape. For cosine vector (4A), the same and similar groups occur evenly within their respective ranges, with same appearing mostly below similar. Different questions are all mostly clustered near the top, with only a few points having a slightly closer distance score (i.e., the outliers we previously mentioned). The graph of the Euclidean data (4B) continues to show the distance measure performing poorly: session distance scores, whether for same, similar or different questions pairs, are in nearly identical ranges.

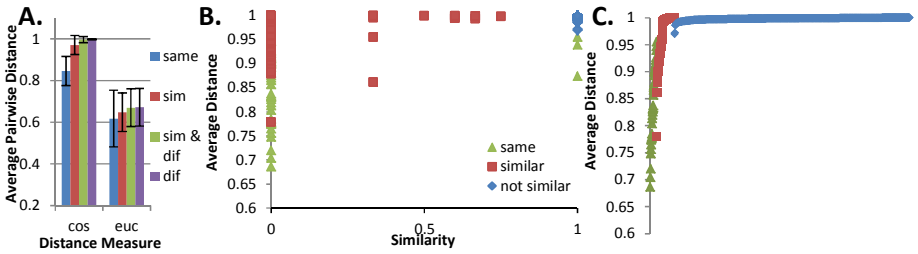


**Fig. 5.** The same graphs from Fig. 1, 2, and 3 using the cancer type distance, $QDist_{ct}$, and cosine vector distance (and Euclidean distance in 4A) with PFISF weighting.

Fig. 5 shows results for the other similarity measure we implemented, $QDist_{ct}$, based on cancer type which we described in Section 3. Here we see three graphs, A, B and C, that correspond to Fig. 2, 3, and 4. Since the results for $QDist_{ct}$ were quite similar to the results for the $QDist_{tfidf}$ with regard to distance measure performance and the effect of the session vector representation, we only show data for PFISF weighting and the cosine vector distance measure except for 5A where we include the Euclidean distance measure.

Fig. 5A shows a similar result to what we saw in Fig. 2; same questions are discernible from dif & sim questions with T-Test scores of 7.42E-16 and 1.01E-09 for cosine vector, respectively. We also see the same difference in performance between the two classes of distance measure.

Fig. 5B compares average session distance per question pair and question distance using cancer types, $QDist_{ct}$. $QDist_{ct}$ is less nuanced than $QDist_{tfidf}$ (Fig. 3); few data points are in the middle meaning that most of the questions either had the same associated cancer types or had no types in common.

When we compare Fig. 5C to 4A, we see that both question distance measures classify the majority of low scoring session distances (sessions with more pages in common) as same or similar. However, we see that $QDist_{ct}$ classifies many more question pairs as similar where the session distance score is large (sessions have less in common). This result explains why there is less distinction between column 2 (sim) and 4 (dif) in 5A than in 2D.

### 4.4 Question difficulty and session vector distance

Our third research question was: how does question difficulty affect the ability of distance measures to discriminate between questions? Recall that we measure a question's difficulty for this analysis as the average number of page views taken for users to answer that question.

Fig. 6 shows average distance between sessions for users answering the same questions plotted against the average length of sessions for those questions. Fig. 6A shows the results for the cosine vector distance measure. Here we see very little correlation between question difficulty and distance between sessions. There is a slight upward slope to the line as the number of page clicks goes up but the $R^2$ value of a line fitted to the data is .22, meaning only 22 percent of the variance can be attributed to it.

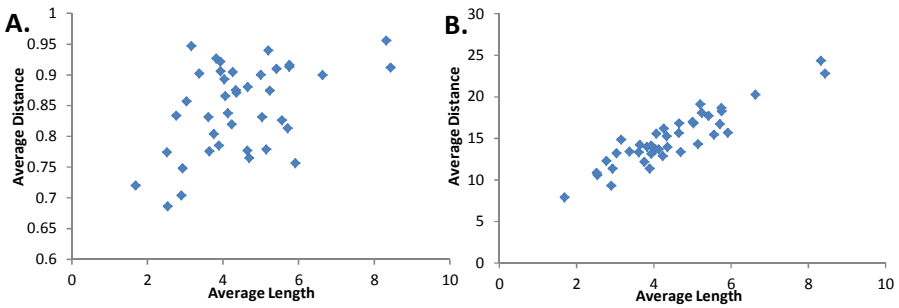


**Fig. 6.** A comparison of average distance and length for same questions for cosine vector, A, and Euclidean, B, distance measures

So for the cosine vector distance measure average sessions length had little effect on the average distance between questions.

For the Euclidean distance measure we see a strong correlation between question difficulty and average distance between sessions. As questions increase in difficulty, they also increase in average distance between sessions. A line fitted to this data had an $R^2$ of .84. This result means that the longer the average users session the farther apart those sessions are, which is surprising to us since we assumed that as sessions got longer there would be more of a chance for sessions to have pages in common which would lead to a lower average distance.

### 4.5 Clustering algorithm comparison

In phase two of our study we used a clustering algorithm to directly compare the performance of distance measures. First, though, we tested 4 clustering algorithms: k-Means, OPTICS, SLINK, and EM to find an algorithm that performed well for our data as described in Section 3.5. Fig. 7 shows the average score of five runs for our five extrinsic cluster evaluation metrics; the error bars represent one standard deviation. We see that k-Means and OPTICS perform the best for precision, F Measure, Fowlkes-Mallows and Rand, with k-Means on top for all but Rand. SLINK is by far the top performer for recall followed by k-Means. The reason SLINK performs so well for recall is that it clustered 1158 of 1223 sessions into one cluster; this produces high recall since most of the sessions answering the same question are in the same cluster, but this is not a good clustering overall as shown by the other metrics.

We tested the significance of the difference between k-Means and the other clustering algorithms for all metrics. Table 2 shows the p-values from those t-test; numbers in bold are significant. We see that k-Means is significantly better than EM and, for all but recall, k-Means is significantly better than SLINK. Comparing OPTICS and k-Means, we see that the only metric for which they are significantly different is recall where k-Means has the edge. We chose to use k-Means as the clustering algorithm for the rest of our analysis, since k-Means offers good overall performance.

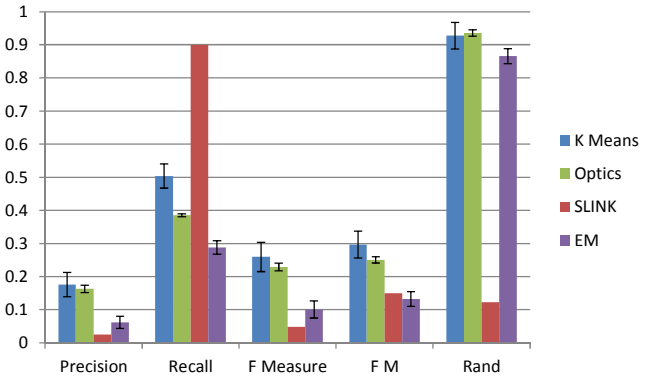


**Fig. 7.** A comparison of extrinsic cluster quality metrics for clusterings of 40 clusters using k-Means, Optics, SLINK and EM and the cosine vector distance measure

**Table 2.** Significance scores (P-values from T-Tests) for extrinsic cluster quality metrics for the differences between k-Means and the other clustering algorithms. Significant values are bold.

| | Precision | Recall | F Measure | Fowlkes-Mallows | Rand |
|---|---|---|---|---|---|
| K-Means/ OPTICS | 0.493526 | **0.001785** | 0.203346 | 0.060922 | 0.247371 |
| K-Means/ EM | **0.000895** | **1.83E-05** | **0.000342** | **0.000166** | **0.02372** |
| K-Means/ SLINK | **0.000791** | **1.71E-05** | **0.000446** | **0.001213** | **2E-08** |

## 4.6 Analysis of distance measures using k-Means clustering algorithm

Fig. 8 shows graphs for 4 of the 5 metrics (F Measure and Fowlkes-Mallows were highly similar) for our 5 distance measures with 40, 80, 120, 160, 200 clusters. Each point represents an average of 5 runs where session order was randomized prior to each run.

Fig. 8A shows precision. For all distance measures, unsurprisingly, as the number of clusters increases so does the precision; more clusters means smaller overall cluster sizes which tends to decrease FPs. The class of distance measures able to differentiate same and different questions in our initial analysis, cosine vector and Tanimoto, perform better than the other distance measures. We also see that Tanimoto performs significantly better than cosine vector for all numbers of clusters. Also interesting to note is that for Tanimoto, the precision gain slows after 120 clusters.

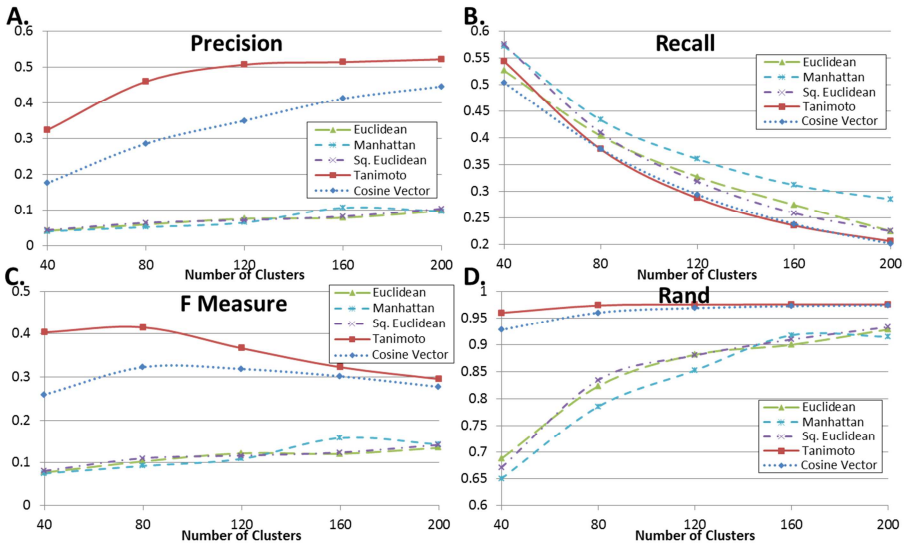


**Fig. 8.** Extrinsic cluster quality metrics, A. precision, B. recall, C. F measure, D. Rand, for clusterings using k-Means and 5 different distance measure for cluster sizes: 40, 80, 120, 160, and 200.

Fig. 8B shows recall. For all measures, unsurprisingly, as the number of clusters increases the recall decreases; more clusters means smaller overall cluster sizes which means more chance for sessions answering the same question to end up in different clusters. For recall, cosine vector and Tanimoto perform worse than the other distance measures. We find that the distance measures with good recall tend to have a single cluster that contains the majority of the points; this is the case for Euclidean, Squared Euclidean, and Manhattan. This single large cluster, while good for recall, is not ideal for our task.

Fig. 8C shows F Measure: essentially an average of precision and recall. Once again we see that cosine vector and Tanimoto perform well with Tanimoto having the edge. Also interesting is that for cosine vector and Tanimoto, the F Measure starts decreasing after 80, which suggest that the optimal k for these distance measures might be between 80 and 120 for our data set. The fact that a cluster size larger than the number of questions we used, 40, seems to be optimal is not surprising when you consider that answers to our questions could appear on multiple pages and that users may have used different methods to find the answer to the same questions (i.e., browse heavy vs. search heavy). Each of these factor could lead to distinct sub clusters within session for the same question.

Fig. 8D shows Rand. Rand represents the proportion of pairs of sessions that were clustered correctly, i.e. true positives and true negatives. Once again we see that cosine vector and Tanimoto perform the best with Tanimoto having the advantage for smaller cluster sizes.

### 4.7 Distance measures

In all of our results, we have found that 2 distance measures perform better than the other three. To explain why these two distance measures perform significantly better, consider the following 3 vectors:

1 0 0 0
1 0 1 1
0 1 0 0

When measuring the distance between vectors of sessions, we believe that sessions that have something in common, vectors 1 and 2, should be considered closer than vectors that have nothing in common, 1 and 3. For cosine vector and Tanimoto distance measures, when two vectors have nothing in common they are always the maximum distance apart, 1. However for Euclidean, Manhattan, and squared Euclidean vectors 1 and 3 are considered closer than vectors 1 and 2. This result leads to short sessions with little or nothing in common being closer than long sessions with more overlap. This property could account for the positive correlation we observe between distance and questions difficulty for the Euclidean distance measure observed in Fig. 6B and is likely a factor in why Euclidean, Manhattan and squared Euclidean perform so poorly for our data.

# 5 Conclusions and Future Work

We found that machine learning distance measures were able to differentiate between sessions from people answering the same question and different questions. Moreover we found that that ability was greatly influenced by the distance measure used, with cosine and Tanimoto performing well and Euclidean, squared Euclidean, and Manhattan performing poorly. We found that the ability to differentiate was marginally affected by session vectors representation and question difficulty. When the two classes of distance measures were used for clustering, our initial results were confirmed. Cosine vector and Tanimoto performed very well; the other three did not perform well.

We observed, for cosine vector and Tanimoto, when we separated similar questions from different questions, regardless of which similarity metric we used, our ability to differentiate same from similar was somewhat less than differentiating same from different plus similar. But it was still possible. Both same and similar questions had a lot of variance in terms of average distance and had a fairly even distribution within that range. What is interesting about this result is that the amount of overlap between user sessions seems dependent on the question itself. It could be the case for same and similar questions, that the ability to tell them apart using our methods is affected by the number of pages available about the question in the collection. This seems to be the case when we consider the outlier we identified in the similar range in Section 4. By looking at outliers, we also observed that there were some limitations of our question distance measure. However, TF-IDF is used extensively in the infor-

mation retrieval community and the limitations are well understood. They did not appear to impact our results other than explaining a few outliers.

We found that when clustering using k-Means, one distance measure, Tanimoto, performed the best. Tanimoto and cosine vector are highly similar measures so it's curious that one performs significantly better than the other. For both measures when sessions have no pages in common the distance is 1 and when they have all pages in common the distance is 0. The difference between the two is that Tanimoto penalizes more for mismatches: where pages are in one sessions but not the other. For example consider the following two vectors:

1 0 0 0
1 1 1 1

The distance between these two vectors is .5 and .75 using cosine vector and Tanimoto, respectively.

It should be noted that, while this study used real questions from actual users and user sessions of real people looking for the answers to those questions, this is still a controlled experiment. This type of question answering behavior is very likely not the only type of behavior on a site. Also, users often do not have well-formulated questions, such as the ones we provided. Further, they may have multiple unrelated information needs in the same session. These challenges should be addressed if one were to use these techniques on actual user session data.

Other interesting work that could be done with our data is to see what effect the correctness of a user's answer has on the distance between sessions. One could also examine whether sessions with the correct answers found the answer on the same page. It is entirely possible if two (or more) pages have the answer for a question, there could be two (or more) paths for finding those pages leading to sessions that are completely different but found the answer to the same question. This possibility could affect our current study by making average pairwise distances between same questions farther apart.

Future work could include enumerating more properties of distance measures for our type of data to further refine our choice of metric or to inform the development of a new one. We may also apply what we have learned about distance measures and representations of session vectors to real log data where user intentions are unknown. Such an effort could be used to enhance search or to suggest pages for users based on historical user access and the perceived information need of the current user. We are also interested in considering the structure of the website and how it relates to user sessions, in addition to just looking at the click log.

Acknowledgements. We acknowledge the support of the Danish Cancer Society and Mr. Tor Øyan, our contact and support from the National Science Foundation, award 0812260. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the NSF. We thank Ms. Tesca Fitzgerald, Ms. Suzanna Kanga, Ms. Flery Decker, and Jonathon Britell, MD, Board Certified Oncologist.

# 6 References


1. D. Beeferman and A. Berger, “Agglomerative clustering of a search engine query log,” in Proceedings of the International Conference on Knowledge Discovery and Data Mining, 2000, pp. 407–416.
2. G. Castellano, A. M. Fanelli, and M. A. Torsello, “Mining usage profiles from access data using fuzzy clustering” in Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization, 2006.
3. E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow, “Using information scent to model user information needs and actions and the Web,” in Proceedings of the SIGCHI conference on Human factors in computing systems, 2001, pp. 490–497.
4. Y. Fu, K. Sandhu, and M. Shih, “Clustering of Web Users Based on Access Patterns,” In Proceedings of the KDD Workshop on Web Mining, 1999.
5. C. Li, “Research on Web Session Clustering,” Journal of Software, vol. 4, no. 5, pp. 460–468, 2009.
6. B. Mobasher, R. Cooley, and J. Srivastava, “Automatic personalization based on Web usage mining,” Communications of the ACM, vol. 43, no. 8, pp. 142–151, 2000.
7. O. Nasraoui, H. Frigui, A. Joshi, and R. Krishnapuram, “Mining Web access logs using relational competitive fuzzy clustering,” in Proceedings of the Eight International Fuzzy Systems Association World Congress, 1999, vol. 1, pp. 195–204.
8. G. Pallis, L. Angelis, and A. Vakali, “Validation and interpretation of Web users’ sessions clusters,” Information Processing & Management, vol. 43, no. 5, pp. 1348–1367, Sep. 2007.
9. W. Wang and O. R. Zaïane, “Clustering web sessions by sequence alignment,” in In Proceedings of the 13th international workshop on database and expert systems applications, 2002, pp. 394–398.
10. T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal. “From user access patterns to dynamic hypertext linking,” in Proceedings of the fifth international World Wide Web conference on Computer networks and ISDN systems, pp. 1007-1014, 1996.
11. N. Jardine, and C. J. van Rijsbergen, “The use of hierarchical archic clustering in information retrieval.” Information Storage and Retrieval, 7, pp. 217-240, 1971.
12. E. M. Voorhees, “The cluster hypothesis revisited,” in Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 188-196, 1985.
13. J. Steinhauer, L.M.L Delcambre, M. Lykke, and M. Ådland, “Do User (Browse and Click) Sessions Relate to Their Questions in a Domain-Specific Collection?” in Research and Advanced Technology for Digital Libraries, vol 8092, pp 96-107, 2013.
14. A. Strehl, J. Ghosh, and R. Mooney, “Impact of similarity measures on web-page clustering,” in *Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, 2000, pp. 58–64.
15. B. J. Jansen, A. Spink, and T. Saracevic, “Real life, real users, and real needs: a study and analysis of user queries on the web,” Information processing & management, vol. 36, no. 2, pp. 207–227, 2000.
16. M. Ageev and Q. G. D. L. E. Agichtein, “Find it if you can: a game for modeling different types of web search success using interaction data,” in Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval, vol. 11, pp. 345–354, 2011.

17. Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. 2002. “Discovery and evaluation of aggregate usage profiles for web personalization,” Data Mining and Knowledge Discovery 6, 1, pp. 61-82, 2002.
18. M. Buhrmester, T. Kwang, and S. D. Gosling, “Amazon’s Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data?,” Perspectives on Psychological Science, vol. 6, no. 1, pp. 3–5, 2011.
19. Mahout, http://mahout.apache.org/
20. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, “The WEKA Data Mining Software: An Update” in SIGKDD Explorations, vol 11, issue 1, 2009.
21. E. Achtert, S. Goldhofer, H. Kriegel, E. Schubert, A. Zimek, “Evaluation of Clusterings – Metrics and Visual Support.” in Proceedings of the 28th International Conference on Data Engineering (ICDE), 2012.