Reda Bouadjenek, Scott Sanner, and Yihao Du
Department of Mechanical & Industrial Engineering
The University of Toronto, Toronto, ON, Canada
26 November 2019

To: Editor, Information Systems

Dear Editor,

We thank the reviewers for their feedback and the time they have taken to read the article and provide us with insightful questions and comments.

We have carefully considered these comments and enclose a list of responses and modifications (highlighted in red in the article) that hopefully have contributed to an improved article.

Reviewer #1

Comment 1. The problem definition is problematic. For example, in Section 2.2, the three variables, I(j), B(j), and S(j) depend on the query and the information element, but the notations give a misunderstanding that they only depend on the information element. A better naming style would be I_q(j), where q denotes the query. Also, giving brief introduction to the querying and results retrieving steps would help readers better understand the problem.
Answer: Thanks for this suggestion, we agree that adding a subscript to refer to the query will improve the notation. Consequently, we have modified the entire notation of the manuscript. We also added a short description of both the query and retrieval steps in Section 3.1.

Comment 2. Figure 1 is confusing: why the position of points changes over subfigures? These points are not placed by longitude and latitude?
Answer: Points are placed by longitude and latitude, however, not all subfigures show the same points. Subfigure (a) shows all results, but Subfigures (b) and (c) show only the *subset* of tweets appearing in the top three clusters for the respective algorithms of (b) and (c). To this end, all points in Subfigures (b) and (c) appear in Subfigure (a).

To clarify, we have modified the caption of Figure 1 to mention that Subfigure (a) shows all tweets while we are only showing the top three clusters and their associated tweets in Subfigures (b) and (c).

Comment 3. The authors adopt F-1 score to measure the coherence of clusters, but a more clear rational should be presented. For example, what does a high F-1 score mean in terms of semantics, spatial and temporal relation?
Answer: Thank you for pointing this out. We agree that more explanation would be helpful here and have modified the article as described below.

Because we are focusing on Boolean retrieval and clusters are the manner by which we return search results, we argue in Section 4.1 that F1-Score on clusters is the only standard Boolean relevance criteria that does not have a pathological solution. We further argue in our revisions of Section 4.1 that F1-Score inherently achieves some reasonable level of cluster "coherence". In short, we argue that coherence corresponds to "tight" constraint settings in all dimensions (spatial, temporal, keyword); while Precision and Recall arguably lead to incoherent clusters (respectively, too small or too large), F1-Score balances these two to return moderately sized clusters. If an F1-Score cluster shrinks unnecessarily, it's Recall component would decrease and make it suboptimal, while if it expands unnecessarily, it's Precision component will decrease and also make it suboptimal. Hence, optimizing clusters for F1-Score does correspond to some *locally optimal* notion of temporal, semantic, and spatial coherence when considering expansions or contractions of the selected cluster.

These claims of increased coherence for relevance-driven clustering are *directly evidenced* in Figure 1, where we see that the relevance-driven clusters of Figure 1(c) are much more "tight" in terms of time span (a few days to a month), spatial extent (well localized), and top keyword content (words are coherent) than K-means shown in Figure 1(b) which has non-localized spatial extent, unnecessarily large time spans of 11 months or more, and incoherent top keywords in a single cluster ("Blizzard", "Earthquake", and "Tornado" together in the right-most cluster). We believe this increased coherence underlies the key reason why our F1-Score optimized clusters lead to faster user identification of event type, location, and date than K-means in our user study (cf. Figure 6).

We have added much of the above discussion to Section 4.1 to help clarify these points for the reader as well as to point out opportunities at the end of Section 4.1 as well as Section 8 for future research augmenting the relevance-focused F1-Score with additional cluster criteria to further enhance word-level, spatial, and temporal coherence.

Comment 4. Are there any parameters (like k in k-means) to control the size and number of clusters? Is it possible that the greedy result consists of only one big cluster or a large number of very small clusters?
Answer: We do allow the number of clusters to display to be specified and make this more clear with some text changes and additions in Section 4.3.

There are certainly some pathological cases where individual clusters may be very large or small, however, as argued in the new additions to Section 4.1, the F1-Score metric tends to avoid these extreme cases since very large clusters sacrifice Precision and very small clusters sacrifice Recall.

Comment 5. What would happen if the greedy clusters are spatially overlapped? How to place and bound these clusters?
Answer: Yes, it may happen as shown in Figure 5 in which two different and relevant events happened at the same general location (though perhaps different times). We address this overlap as a UI issue by providing the user with buttons to show/hide individual clusters as shown in Figure 5; we clarify how spatial overlap is addressed in the revised caption of Figure 5.

Reviewer #2

Comment 6. The abstract is rather long.
Answer: We have shortened the abstract to state briefly the purpose of the research, the motivation, the problem addressed, the principal results and major conclusions.

Comment 7. The introduction is much about a technical/practical problem, but too little about a knowledge gap. Perhaps take a look at the work on "faceted search" or "faceted classification", where the idea of using time and space to classify information items dates back to S. R. Ranganathan. This idea has been long pursuit, especially in the HCIR community.
Answer: Thanks for the suggestion. Indeed, the reviewer can probably tell that we (the authors) viewed this work from the perspective of improved search-driven visual clustering. But we completely agree that another angle on this work (with its own literature) would be from the perspective of automating filtered and faceted search. We have thus modified the Abstract and Introduction and Related Work (Section 2) to discuss this perspective with added citations for modern information retrieval views of filtered and faceted search as summarized in works by Tunkelang, Hearst, and co-authors.

We remark that our baseline method in Figure 1(a) represents a manually driven multiple filter search approach where the user must manually specify the time window sliders (shown in Figure 5), pan and zoom level to select a geographical spatial area, and keyword search (also shown in the evaluation interface of Figure 5) to drive the search process. From this perspective, we now explain that our goal in this article can be viewed as (approximately) optimally extracting filter settings (per cluster) according to relevance-driven criteria that can substantially reduce manual filter tuning efforts. We also now explicitly note the relation of our baseline method to filtered search when we introduce this method in Section 7.

Comment 8. I am a bit in doubt about the novelty of the presented probabilistic re-interpretation to precision/recall, but I cannot conclude on this with certainty. Perhaps take a look at "A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation" by Cyril Goutte and compare your approach to theirs.
Answer: First we want to be careful not to claim here that our probabilistic re-interpretation of Precision/Recall/F1-Score is novel and we have modified the text to avoid this implication; while we explain below why Goutte and Gaussier's work is different, we only wish to argue that we need to take an *expectation* of F1-Score for our task since relevance is uncertain.

We thank the reviewer for the suggested reference "A Probabilistic Interpretation of Precision, Recall and F-score, with Implication for Evaluation" by Goutte and Gaussier, which we have now cited and discussed in Section 4.1. However, we remark that while we analyze the expectation of F1-Score under a per-document relevance probability estimate provided by a probabilistic information retrieval system (e.g., the language modeling scoring function we use) that varies for each document (cf. Equation 5), Goutte and Gaussier instead focus on a Bayesian posterior analysis of the underlying generative parameters for precision and recall (that are considered to be the same for all documents drawn from an i.i.d. distribution). In short, their focus is on obtaining a posterior estimate of p and r (used to compute F1-Score) given *ground truth* for relevance while our focus is on estimating the F1-Score from uncertainty estimates of relevance per documents in the absence of ground truth. Accordingly, the analyses and purposes are very different.

Comment 9. I feel the text is rather repetitive. Check for possible repetition of phrases/statements.
Answer: We have removed a number of redundant phrases in the Introduction, Problem Definition, and Algorithm Description.

Comment 10. How does your clustering method compare to (probabilistic) topic modelling methods, such as LSA?
Answer: Clustering and topic modeling methods are very different in nature. While clustering attempts to aggregate similar documents with the assumption that each document belongs to one cluster, topic modeling methods such as (probabilistic) LSA assume that a document is composed of multiple topics and try to estimate the degree or probability of relevance that a document has to each topic. While one might imagine a *much more complex* version of this article that somehow attempts to optimize topic models for spatio-temporal visual display, such an extension is a number of technical and computational leaps beyond the current article (highly effective spatio-temporal extensions of topic models, novel methods to visually display topics as opposed to clusters, extensions to topic modeling that explicitly consider relevance-based optimization criteria, and novel computational methods to effectively optimize within this extended framework). Nonetheless, it is an interesting suggestion and we have added a short discussion of this possibility in our future work discussion of Section 8.

Comment 11. I am rather confused about the signals used for re-ranking, in particular w.r.t 'relevance'. You write that you optimize your clusters for relevance, and you mention language models, but also a term-weighting/pruning method. In general, I think the article would improve with a more high-level introduction of the signals used for re-ranking and how they are obtained.
Answer: We believe the key point of confusion may arise in our introduction of language model scoring in Section 7.1 and how it interacts with term selection (pruning) in cluster extraction so we add a short discussion in this section regarding the different ways our framework uses them. To explain briefly here, there seems to be some confusion between cluster filter criteria, relevance scoring, and the evaluation framework (the only place we mention "weighting", though we note that we never mention any "term-weighting"). A cluster is defined by a spatial bounding box, time interval, and the set of words included/excluded. Separately, we need a probability that each document is relevant to a query and this is obtained from a language model scoring function. The evaluation framework uses a lambda

weight parameter to modulate the signal to noise ratio in our experimental design, but we remark that this does not pertain to any sort of term-weighting.

Comment 12. Please specify the grid search that you perform for the K-means distance metric. The resulting distance equation is rather counter-intuitive to me, as it focuses too much on one variable which makes it rather indifferent from a non-combined distance metric based on this variable alone. I wonder how this affected the outcome of the user study. Also think about Pareto-optimized or relevance-first type of ranking mechanisms, which take a more balanced or stratified approach to calculating relevance (and similarly, distances).
Answer: We have clarified the methodology and range/increments for grid search tuning of the distance metric for K-means in Section 5.2. We cannot definitively comment on how changes in these parameters would affect the user experiments though we do note that other values (e.g., balanced weights) led to clusters with more extreme dimensions that cluttered the user interface.

We appreciate the suggestion of Pareto-style approaches, but we must admit that we do not fully understand the reviewer's suggestion here and exactly how it would be implemented, hence we did not want to include a discussion based on a possible misunderstanding from our end. We would almost suggest that our "distance-free" method of relevance-driven clustering algorithm (BPS) might actually coincide with this suggestion in that our clusters expand to include relevant information and contract to exclude irrelevant information and hence do not require explicit notions of distance but only the relevance-impact of expansions and contractions in each dimension. We remark that it was precisely the difficulty of defining a relevance-first distance metric for K-means that led us to the proposed relevance-driven approach to clustering that does not inherently require tuning of dimensional trade-offs.

Comment 13. For the user study, you use a rather extensive questionnaire with a lot of factors. Please argue for which factors are relevant to your study and why these would be affected by the chosen algorithms. Also, specify the hypotheses more clearly (and correctly): What effect do which algorithms have on which constructs?
Answer: This is a good point and we believe these changes will clarify the NASA-TLX analysis. To address this request, we have added the following two changes to Section 7.4.1:

"Overall, we hypothesize that the three most important factors for this visual search task are temporal demand (reduced time to complete the search task owing to better clusters), mental demand (the ability to focus analysis at the cluster level of abstraction as opposed to the tweet level), and effort (reduced effort to analyze clusters due to clear keyword summaries). We conjecture that all of these task load reductions would follow from the increased coherence of the BPS relevance-driven cluster extraction compared to unsupervised K-means clustering and the lack of any clustering in the filtered search Baseline."

and

"We note that participants overall perceived the BPS algorithm to be more effective at helping them complete their search task in comparison to using K-means or the Baseline. Considering each of the three aforementioned key factors (TD, MD, EF) deemed most relevant to the innovations of the BPS clustering algorithm, we remark that BPS recorded the lowest median load on all three factors with K-means somewhat behind in second place (indicating that some form of clustering still aided the visual search task) and the filtered search non-clustering Baseline further afield with the highest loads."

Comment 14. Is the used baseline strong enough? What are baselines in similar studies of geographic/faceted search?
Answer: For the offline experiment, we used an *optimal* Mixed Integer Linear Program (MILP) formulation to benchmark the performance of our algorithm on moderately sized datasets. So given that the MILP formulation provides the optimal solution, we believe that there is no better baseline here for comparison.

For the user study, one of our baselines is precisely a manually-driven multiple filter search approach and the other method is the most commonly used clustering method used in practice (K-means) that has historically been used extensively in the past for visual search clustering tasks. Adding another baseline would have made the trial length unreasonably long (and exhausting) and increased the number of users and dataset scenarios required to experiment with all permutations required to properly randomize the evaluation as currently done. Hence, with the restriction to evaluating two baselines, we believe we have chosen the two most important baselines relevant to existing methods for the visual search task evaluated in our user study.

We have added clarifying discussion relating to our rationale for baseline selection in the discussion of Section 7.2.

Comment 15. Please move your related work to the beginning of the paper, as to make it part of the whole and use it to derive certain outstanding 'areas for improvement'. Also, be careful not to have too much 'dated' literature in this section.
Answer: We have moved the related work earlier in the paper to the new Section 2. We are aware that we have many older references, but we have done an extensive literature survey and believe it is critical and necessary to cite the earliest work that addressed key questions in each area of related work that we cover.

Comment 16. Consider defining sections 'Study 1' and 'Study 2' in order to more clearly distinguish the different evaluations performed. Also, try to "sell" this dual approach more.
Answer: Thanks for the suggestion. We have renamed the sections Study 1 and Study 2 in order to more clearly distinguish the different evaluations we performed. We have also justified this dual evaluation approach by explaining in Section 5.1 how the offline evaluation methodology allows us to literally run 1000s of trials to understand how our greedy algorithm compares to other clustering methods (including the optimal MILP solution) as we vary properties of the data and relevance score noise. Leading up to Section 6, we have now also explained that, on the other hand, the user study allows us to literally put the human-in-the-loop with a real search task to determine if relevance-driven

clustering achieves its intended end-goal of improved visual search performance in an end-to-end comparative evaluation with competing alternatives.

Comment 17. Considering co-publishing your Twitter dataset (Table 1) including its ground truth, fx in a DataVerse repository with a DOI attached.
Answer: Unfortunately, Twitter's developer policy places a number of restrictions on the use of their API and the data obtained from it. Some of the key restrictions we encounter are the following:

1. In the Restrictions on Use of Licensed Materials (II.C), Twitter states that one can only use geographic data to identify the location from which a tweet was made and not for any other purpose. Elsewhere they place further restrictions on storage of tweets. They reiterate this in section B.9
2. Section F.2 "Be a Good Partner to Twitter" is a key issue for data redistribution, as here one is only allowed to redistribute the ID for a tweet (no message content, user data, etc). There are further restrictions on how many tweet IDs you can publish per user, per day.
3. In the Ownership and Feedback section, Twitter makes it clear that they can revoke data publishing rights at any time in the future, which is a further concern if data is published with a permanent DOI.

Hence, as we understand it, it would not be possible to publicly post our data in a way that would be of much use to our readers. That said, in the interest of reproducibility and science, we would privately share the dataset with anyone who emails us (the paper authors) to request it. We're not sure however that we can write this officially in the article itself.

Comment 18. Separate the tasks in the user study between "hard" and "easy" tasks. This is common in IR, where hard tasks have only few documents in the collection that contain the answer.
Answer: We remark that our existing user study used uniformly difficult tasks since we had to randomize both order and assignment of tasks to the three different search interfaces in order to avoid introducing bias in the evaluation. While we acknowledge that creating easy and hard tasks would be interesting in a subsequent extended user study, it would have far exceeded our user study time and grant-supported monetary budget to have attempted to do this in the present article.

Thank you for your consideration. We hope that the answers to the above comments will clarify the concerns the reviewers have raised while improving the article in places that the comments necessitated revisions (as noted above). We remain available for any inquiry that you may have. We look forward to hearing from you.

Sincerely,
Reda, Scott, and Yihao

Attachment:

- Article Manuscript