Reda Bouadjenek, Scott Sanner, and Yihao Du
Department of Mechanical & Industrial Engineering
The University of Toronto, Toronto, ON, Canada
22 March 2020

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.

We also note that to improve clarity of the three contributed algorithms and their relation to each other, we have renamed our algorithms as: (1) RadiCAl-MILP, (2) RadiCAl-Greedy, and (3) RadiCAl-BPS, where "RadiCAl" is the acronym of "Relevance-driven Clustering Algorithm". This has been modified throughout the paper.

---

# Reviewer #1

**Comment 1.** The authors have addressed most of the issues raised in my previous review except for the second weak point "the clustering approach lacks novelty". Other parts of the paper look good to me now.

Answer: We would agree in general that clustering has been previously applied in visual search interface tasks, but we respectfully disagree that the clustering approach proposed in this paper is not novel. In short, almost all clustering is inherently unsupervised and requires the definition of a distance metric as in our K-means baseline. In contrast, we have introduced a novel formulation of relevance-driven clustering for visual search as an optimization problem given a probabilistic measure of relevance for search results. We have motivated and derived *expected* F1-Score as an objective criteria for relevance-driven cluster optimization. Furthermore, we have shown that this *expected F1-Score* optimization problem can be reduced to an optimal Mixed Integer Linear Program (MILP) solution which we use to benchmark the greedy algorithms we propose.

In short, we are not aware of any previous approach that ties together visual search interfaces and information retrieval in a clustering algorithm as proposed in this article and we believe that this underscores it’s novelty. We note that we have reinforced some aspects of this novelty description while addressing Reviewer #2’s comments below, which should also help address this comment.

# Reviewer #2

**Comment 2.** I keep finding it a bit irrelevant that the abstract mentions all different parts of the questionnaires used. Try to either indicate why it's relevant to look at these parts, or just mention the effects measured for the whole questionnaire.

Answer: Thanks for pointing that out. We have streamlined this explanation by now ending on the following sentence summarizing our key results: "these higher-relevance clusters that have been optimized w.r.t. user interface display constraints result in faster search task completion with higher accuracy while requiring a minimum workload leading to high effectiveness, efficiency, and user satisfaction among alternatives."

**Comment 3.** The introduction and problem statement are rather short. I would recommend stating the problem statement more clearly in the introduction. What are the current interaction issues with spatial-temporal-content visualization? And with ranking? What other attempts to fix these issues have been done, and where were they lacking? What are the types of user tasks and technical challenges that you address, and why are they common/significant?

Answer: We have previously motivated our work w.r.t. shortcomings of clustering for visual search interfaces (which almost categorically use K-means and has many deficiencies for visual search clustering as outlined in the introduction), but we have widened this motivation to a broader information retrieval ranking context on pages 2-3:

1. We start by giving a real example related to the task of searching a multiyear Twitter corpus for content related to natural disasters.
2. Next, we mention that this problem can be addressed using a conventional information retrieval list ranking approach based only on the textual content of the tweets.
   - **Critique:** we are mentioning two facts here:
     - i. This would flood the user with many elements that are matching his query, thus, preventing him/her from carrying out an efficient investigation.
     - ii. This won't fully take advantage of the spatio-temporal information associated with the tweets.
3. A more sophisticated approach that we describe next would be to use a typical spatial-temporal-content visualization approach that would provide all matching tweets in a map-based display (Figure 1(a)). This will certainly improve the presentation of the search results compared to the list approach
   - **Critique:** Although this will certainly improve the presentation of the search results compared to the list approach, we argue that it will also take the user a large amount of time to sift through.
4. An attempt that can be used to improve the previous strategy would be to use filtering and faceted search. This later can be used to help the user manually narrow the large set of tweets using filter settings defined for each tweet aspect (location, posting time, keywords).

- **Critique:** We argue that a large amount of effort is still required on behalf of the user to manually read through results and adjust filters appropriately.

5. To improve the previous interface and ease the task of browsing search results, we can cluster results display like that shown in Figure 1 (b). This way, we can restrict the displayed information such that similar tweets appear together. We mention that most existing work on aggregation for visual search that has sought to exploit the cluster hypothesis has focused on K-means and related unsupervised clustering methods.
    - **Critique:** we are mentioning three facts here:
        - i. A conventional clustering approach does not necessarily guarantee that clusters of matching search results are highly relevant.
        - ii. The use of clustering algorithms such as K-means requires the design of a complex distance metric; such ad-hoc metric specifications do not necessarily guarantee the coherence of clusters from a visual, temporal and keyword content perspective.
        - iii. Such an algorithm requires intensive tuning in the distance metric. For instance, space is often measured by Euclidean distance while keyword content is often measured by cosine distance and both of these distances need to be combined into a single distance metric for K-means, which is extremely non-trivial.
6. To deal with the above problem, we argue that a clustering algorithm that is actively aware of the relevance probability of each tweet to the search query could automatically generate highly relevant clusters covering a large fraction of relevant content without requiring an explicit distance metric.
7. Finally, we summarize the technical challenges that we address in this work by mentioning that our contributions in this article can be viewed as (a) avoiding the distance-tuning complications of unsupervised methods like K-means, while (b) automatically extracting filter settings per cluster by optimizing relevance-driven criteria that reduce the need for manual tuning.

We believe that this line of argument as laid out in the Introduction addresses all of the reviewer's criteria regarding contextualizing the work and identification of challenges that we address. As remarked in footnote 1, we believe that the discussion and challenges are common to spatio-temporal visual content user search tasks with a large volume of matching search results; as noted, our running example of Twitter simply provides an exemplar of such a use case familiar to both readers and our experimental subjects.

Comment 4. Related to that I would also highlight more the technical and user challenges that you address. One of the main features of your work is that you address (and evaluate!) both sides.

Answer: We believe that now, our motivation/problem definition in the Introduction (pages 2-4, changes highlighted in red), as detailed above highlights clearly the technical and user challenges that we address in this work. Namely we outline the need for a more visual search problem-specific objective and the need to optimize if efficiently. We also mention that we need to conduct both an offline evaluation to understand the trade-offs with our algorithmic contributions and a user study evaluation to provide both a quantitative and qualitative overview of the benefits of our contributions.

**Comment 5.** I feel there is often repetition in the beginning. Mainly between the abstract, contributions, and related work. Often, repetition is a sign that the structure you're using is not working that well. I'd recommend to reconsider more carefully what should (not) be in the first three sections.

Answer: The points are well-taken though we respectfully differ in opinion on some details and hence provide a multi-part answer here.

We agree that the Introduction had redundant components and we have done our best to remove such content and streamline the explanations as indicated by changes on pages 2-4 (highlighted in red). We especially note that the discussion of contributions has been reorganized starting in the bottom half of page 3 and continuing on to page 4 to prevent repetitive content and better group contributions according to common function. Some redundancy between the Introduction and Abstract is expected since the Abstract is a high-level summary of the entire article and the Introduction seeks to lay out the key arguments and contributions of the article.

Specifically regarding the Abstract, we note that it contains roughly 200 words, which is a typical abstract length. We believe that the function of an abstract is to allow readers to judge whether or not the paper is of relevance to them. It should therefore be a concise summary of the paper's aims, scope, and conclusions. *So, we believe that having an overlap with the other sections of the paper is OK here.* We remark that we have organized and slightly revised our Abstract in a standard format [1] as follows:

1. A general statement introducing the broad research area of the particular topic being investigated as well as the specific problem to be solved.
   - *"Search results of spatio-temporal data are often displayed on a map, but when the number of matching search results is large, it can be time-consuming to individually examine all results, even when using methods such as filtered search to narrow the content focus."*
2. A review of existing or standard solutions to this problem and their limitations.
   - *"This suggests the need to aggregate results via a clustering method. However, standard unsupervised clustering algorithms like K-means (i) ignore relevance scores that can help with the extraction of highly relevant clusters, and (ii) do not necessarily optimize search results for purposes of visual presentation."*
3. An outline of the proposed new solution.
   - *"In this article, we address both deficiencies by framing the clustering problem for search-driven user interfaces in a novel optimization framework that (i) aims to maximize the relevance of aggregated content according to cluster-based extensions of standard information retrieval metrics and (ii) defines clusters via constraints that naturally reflect interface-driven desiderata of spatial, temporal, and keyword coherence that do not require complex ad-hoc distance metric specifications as in K-means."*
4. A summary of how the solution was evaluated and what the outcomes of the evaluation were.
   - *"After comparatively benchmarking algorithmic variants of our proposed approach -- RadiCAl -- in offline experiments, we undertake a user study with 24 subjects to evaluate whether RadiCAl improves human performance on visual search tasks in comparison to*

> *K-means clustering and a filtered search baseline. Our results show that (a) a binary partitioning search (BPS) variant of RadiCAl is fast, near-optimal, and extracts higher-relevance clusters than K-means, and (b) clusters optimized via RadiCAl result in result in these higher-relevance clusters that have been optimized w.r.t. user interface display constraints result in faster search task completion with higher accuracy while requiring a minimum workload leading to high effectiveness, efficiency, and user satisfaction among alternatives."*

**Reference:**

[1] Zobel, Justin. Writing for computer science. Vol. 8. New York NY: Springer, 2004.

On the topic of Related Work redundancy, we only repeat concepts from the Abstract and Introduction when required to differentiate our contributions and requirements from previous work. We do not believe any additional content can be removed while maintaining the same key points of the Related Work discussion.

**Comment 6.** In particular, I feel the contributions are too much detailed/elaborate now. Consider limiting the description to how you address the problems you identify.

Answer: Thank you for this suggestion. Indeed, we recognize that we previously had a long summary of contributions. We have significantly shortened and reorganized our contribution bullet points into three coherent contributions. We have removed some details but retained other details that we believe are critical for outlining the core argument and flow of contributions in the paper as they relate to the previously identified challenges (as suggested by the reviewer).

**Comment 7.** In the related work section, you describe other techniques and then typically conclude they are different. What would be more interesting is to discuss their shortcomings with respect to your goals. Thus concluding "there exists a need for". I'd rather see the related work conclude with several shortcomings, which you then address/solve in a subsequent section.

Answer: We agree that the differences between previous work and our contributions were not previously framed *explicitly* in terms of our goals so we have reworked the related work conclusions of each subsection to help clarify these points:

1. Spatio-temporal clustering: for which we discuss the shortcomings with respect to our goals by saying:
   - *"We note that all of these clustering methods fail to jointly address our goals for visual search clustering as stated in the introduction. Namely, these methods (i) ignore relevance signals (beyond the initial search), (ii) ignore the \emph{joint combination} of spatial, temporal, and keyword constraints, and/or (iii) ignore definitions of clusters that pertain to their presentation in a visual display medium, all of which are key jointly intertwined contributions of the clustering approach proposed in this work."*
2. Clustering in IR: for which we discuss the shortcomings with respect to our goals by saying:

- *"When considering our search-based clustering needs in this article, all of these methods (i) do not explicitly use the relevance signal during cluster optimization to ensure high relevance of extracted clusters, and (ii) do not specifically formulate clusters in terms of spatial, temporal, and content constraints to ensure coherence and succinct presentation in a visual search display. Both of these requirements are addressed in our proposed contributions."*

3. Filtering and Faceted Search in IR: for which we discuss the shortcomings with respect to our goals by saying:
   - *"In this work, we have an explicit query that drives construction of our filters. Further, we directly optimize our filter settings w.r.t. a relevance-based objective to maximize the expected F1-Score given a probabilistic measure of relevance. While these techniques may be used to extend work in information filtering, no existing information filtering work performs the same relevance-driven cluster (or filter) optimization that we propose in this work."* Furthermore, we are argue that "*While our methods arguably build on ideas in multiple filter search and we compare to a filtered search baseline, the key distinction is that existing filtered search has focused on the user interface design and user studies, whereas our work focuses explicitly on automatically extracting clusters (where an individual cluster corresponds to a setting of multiple filters) to maximize relevance-driven optimality criteria."*
4. Explicit Optimization of IR Metrics: for which we discuss the shortcomings with respect to our goals by saying:
   - *"no other existing work has proposed an expected F1-Score relevance-driven optimization approach to clustering as we do here"*. Moreover, we argue that "*L2R (Learning to Rank ) cannot be applied in our cluster optimization problem because we do not have labeled data to train with --- while we have a relevance signal, true cluster labels are not known for any data. Moreover, the task we address is to find cluster settings that optimize an expected Boolean metric of expected F1-Score -- not to optimize metrics for ranking."*

**Comment 8.** Also, be careful about repetition here.

Answer: We have tried to be careful in this latest revision to remove repeated content.

**Comment 9.** The problem definition section does not really state a problem definition: What is the problem with existing clustering methods? With visualization?

Answer: We have significantly reformulated the problem definition to clarify the technical definitions and requirements and to refer specifically to design requirements for both clustering and visual criteria that were motivated in the introduction. All changes to this effect are highlighted in red on pages 6-7.

**Comment 10.** Consider merging the background with related work section.

Answer: We first remark that our "background" is actually more of a framework and notation definitoin section so we have renamed this section "Framework and notation". We considered the request to merge this content into the Related Work, but ultimately feel that it would hurt the article more than it would help for the following key reasons:

(1) We believe that it is often useful to have a Related Work section that stands alone, such that the reader can read or skip it depending on his/her interest which is, in particular, useful for the reading of this paper. Also, as the "Framework and background" section contains critical information for the paper that cannot be skipped, we believe it's better to have it in a separate section.

(2) Because the contributions of this article are so different from Related Work (as outlined in Related Work), there is no obvious unified formalism for discussing all Related Work and the contributions of this article. To this extent, we believe that our "Framework and background" is a distinct and standalone section from Related Work.

**Comment 11.** In your research problem, you refer to "efficiently", "high-relevance", "coherent", and "information need". Make sure each of these concepts are introduced properly before using them in a research question. Also, the research problem should be at the end of the introduction, rather than in the background section.

Answer: Thanks for these suggestions. We have now provided accompanying definitions for all of the terms noted in the problem definition of pages 6-7:

1. For relevance and information need: "*In this paper, we adopt an IR interpretation of relevance, which refers to how well a document or a cluster meets the user's information need. An information need is defined as the information that satisfies a conscious or unconscious need of the user and is formally expressed by the user's keyword-based query [69]. In this paper we will specifically use a language model definition of probabilistic relevance w.r.t. a user's Boolean "or" query [70]..*"
2. For coherent: "*We define a coherent cluster as a group of elements that are topically similar to each other (i.e., similar text content) and that are similar in both their time and space dimensions. We remark that coherency in the spatial dimension forms a key requirement for clusters that can be easily (i.e., compactly) visualized.*"
3. Efficiently: "*How can we efficiently optimize this objective for use in real-time visual search on large corpora?* " We remark that since the clustering problem we are optimizing is NP-Hard, it would be difficult to restrict ourselves to provably polynomial-time clustering algorithms, hence we simply define efficiency in terms of our real-time requirements for large corpora.

Regarding the problem definition, we note that the high-level description is in the introduction, but we wait until the Problem Definition section to formally define all details. As for Related Work, previous reviews have asked us to move it forward immediately after the Introduction -- because the Problem Definition section is highly technical and closely linked to the rest of the contributions that follow, we are reluctant to further reorganize the paper structure. In fact, having a high-level Problem Description in the introduction followed by a discussion of why Related Work does not address this problem, followed by a highly technical definition that leads into subsequent solutions seems to us to be the "right" ordering of content in terms of accessibility and flow for the reader.

**Comment 12.** Refrain from using "pathological", that sounds too negative. It is well-known that precision and recall are "at odds" with each other. You can optimize for one, for the other, but the challenge is to optimize for both.

Answer: Sure, we have softened our proposal of F1-Score as follows: "(expected) F1-Score of clusters is the only standard Boolean relevance criteria that balances all of our cluster desiderata". We later refer to unintended solutions (all content vs. a singleton) as "undesired" solutions. We hope this reads in a more positive light.

**Comment 13.** I feel its quite genius to use EF1 to optimize clusters, given that it strikes a balance in cluster size. Perhaps add a short note on how common/novel it is to use F1 score to evaluate (supervised) clustering techniques, how it compares to other cluster optimization metrics, and what the constraints of this metric are (fx in terms of the distribution of Sj).
Answer: We

**Comment 14.** Similarly, you added a note on the other paper that takes a probabilistic approach to precision and recall. Perhaps shorten this note.
Answer: This reference has been suggested by reviewer 1 who asked us to take a look at the paper and compare our approach to theirs. We have stated that we do not claim 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 mentioned that e 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 15.** I'm curious about how your pruning algorithm compares to several existing feature selection methods. Take, for example, "backward elimination" for regression modelling. Rather than F1, this uses a "goodness of fit" measure to see which features to eliminate. Or take, for example, the "information gain" and conditional entropy that we can use to eliminate terms that are not discriminatory two (known/supervised) classes. Note that I'm curious, but that it is fully up to you to decide whether.
Answer: We have already considered the option of using mutual information to extract the most informative terms, and then greedily prune them to build clusters. However, this method was costly as it needs to compute mutual information for each word in the cluster w.r.t. the scores and the performance were statistically comparable to those obtained simply by using frequency-based pruning. So we had to abandon this approach.

**Comment 16.** Could/will you release your ground truth data set? Sounds like a very valuable data set for the community to base some of the future work that you suggest on.

Answer: In the interest of reproducibility and science, we share the dataset. However, we point out the fact that the 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.

The dataset can now be accessed on the Github repository of this project:
https://github.com/D3Mlab/viz-ir/tree/master/twitter_dataset

**Comment 17.** Could you make it more explicit how you calculate F1 score for your ground truth? Do you, for example, base the ground truth Bq(j) on the relevance of a document to a particular cluster from Table 1?

Answer: We want to clarify, recall and emphasize a critical point here! Clusters are built by optimizing EF1 that is based on the relevance scores of the elements obtained at query time (S_q(j)), whereas we evaluate the quality of the clusters generated using F1 that is based on the ground truth (B_q(j)), a standard procedure in any conventional IR task. We have clarified this at the beginning of Section 6.

**Comment 18.** Could you make the procedure followed in the offline evaluation more clear? You mention right in the beginning of Section 6 "thousands of trials", but it'd help to have a clear overview of how these trials are formed.

Answer: There may be a misunderstanding here. By thousands of trials, we mean thousands of participants that would assess performance based on the size of the dataset, the noise injected, and the number of relevant elements retrieved by a query and that for four algorithms we evaluated in this offline evaluation. This would exhaust us in terms of time and budget allocated for this project.

**Comment 19.** Were there harder vs. easier to solve tasks / topics in the online evaluation? Can you differentiate between them?

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.

Moreover, the feedback we got from participants didn't mention the fact that there were easy and hard topics, which indicates that the difficulty was uniformly distributed over our tasks.

Comment 20. Can you be more explicit about the instructions the participants received. Perhaps also compare your tasks to what is common in interactive IR experiments (fx see Kelly (2009) and Borlund (2003)).

Answer: Borlund (2003) and Kelly (2009) proposed a framework for evaluating interactive IR systems by placing evaluators into task scenarios. In particular, Borlund provides advice for constructing simulated scenarios and tasks, including creating a situation that is relevant to the subjects and providing details to engage the subject in the search in a way which is as close as possible to actual information searching and IR processes. Borlund stated that this can be achieved by the use of the following key components:

1. the involvement of potential users as test persons (called also participants),
2. the application of individual and potentially dynamic information need interpretations,
3. the assignment of multidimensional and dynamic relevance assessment.

Although our evaluation approach can be seen as an instantiation of Borlund's framework, we didn't bother to use (2) as we wanted to stay in a relatively controlled evaluation environment. Moreover, the fact that participants were asked to search for events represented by a collection of tweets rather than evaluating the relevance of individual documents, the use of a binary-based relevance measure (recall) was more than enough for our task, thus, we didn't consider (3) either.

Specifically, before starting the experiment, basic instructions were given to each participant among which not make a break, not to use their phone, and not to use social media such that to still focused on the task and perform a fair evaluation. We have collected basic information related to each participant including the English level, the education degree and the extent to which they are aware of natural disasters. The purpose of that entrance survey was to make sure that there is no bias among the participants. Before starting the formal experiment, each participant was shown a training video describing the visual search interface, how to use it and how to interact with the clusters. Then, each user was tested on his ability to find each of three natural disasters from Table 1 using both the **Baseline** non-clustering interface as well as the **K-means** clustering interface in two training trials. It was important to us to make sure the users understood the global task before starting the formal experiment, thus the need for a training phase.

We have clarified this at the end of Section 7.2.

Comment 21. Try turning the conclusion section into a discussion. It is rather short and mainly a summary like this. Fx. address limitations about the study's conclusions. Were there any experimenter effects? Any confounding explanation? Etc. Or, in which cases/queries/tasks would this technique (not) work / be helpful? And, on which other dimensions could/should we cluster? What if we have many dimensions? Are there any theoretical implication w.r.t. to multi-objective optimization (Deb, 2014)? Or any practical implications w.r.t. reranking in IR (Ghorab, 2012)?

Answer: We have rearranged the sections by adding a discussion section (8) and a conclusion section. The discussion section now contains a summary of the results, the main limitations of this research and the possible future work that we intend to investigate.

In the limitation section, we have addressed most of the questions raised by the review.

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