Proceedings of the ASME 2016 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/DTM 2016
August 21-24, 2016, Charlotte, USA

IDETC2016-59926

DISCOVERING DIVERSE, HIGH QUALITY DESIGN IDEAS FROM A LARGE CORPUS

Faez Ahmed*
Dept. of Mechanical Engineering
University of Maryland
College Park, Maryland 20742
Email: faez00@umd.edu

Mark Fuge
Dept. of Mechanical Engineering
University of Maryland
College Park, Maryland 20742
Email: fuge@umd.edu

Lev D. Gorbunov
Dept. of Electrical and
Computer Engineering
University of Maryland
College Park, Maryland 20742
Email: levdavid@umd.edu

ABSTRACT

This paper describes how to select diverse, high quality, representative ideas when the number of ideas grow beyond what a person can easily organize. When designers have a large number of ideas, it becomes prohibitively difficult for them to explore the scope of those ideas and find inspiration. We propose a computational method to recommend a diverse set of representative and high quality design ideas and demonstrate the results for design challenges on OpenIDEO—a web-based online design community. Diversity of these ideas is defined using topic modeling to identify latent concepts within the text while the quality is measured from user feedback. Multi-objective optimization then trades off quality and diversity of ideas. The results show that our approach attains a diverse set of high quality ideas and that the proposed method is applicable to multiple domains.

INTRODUCTION

When generating creative designs, both practicing designers and researchers agree: "If you want to have good ideas, you must have many ideas." [1] Why? Because having many ideas helps a designer—or a team of designers—explore a design space and find new inspiration from unlikely places.

But is more always better? When do 'many ideas' turn into 'too many ideas'? This paper explores what happens when the number of ideas grows beyond what any one designer can ever hope to use for inspiration; when more ideas become a barrier to inspiration, rather than a strength. It answers the question: "How do we provide inspiration that retains the good parts of having many ideas—increased diversity and quality—without overwhelming designers?" The paper proposes a practical, computational method of making sense of many ideas and then recommending manageable, diverse, and high-quality sub-sets of inspiration to designers.

Other researchers have tackled parts of this problem, primarily around evaluating creative sets of ideas or leveraging large design databases to inspire designers. In the former, Shah et. al. [2] provide metrics for ideation effectiveness where the main measures for goodness of a design method are how they expand the design space and how well they explore it. Kudrowitz and Wallace [3] suggest metrics to be used to narrow down a large collection of product ideas while Green *et al.* [4] propose methods for creativity evaluation through crowd sourcing. In the latter, researchers focused on inspiring designers [5] and inspiring creativity [6]. Outside of design, researchers have tackled this same problem, but for applications like web-page ranking or document summarization. Researchers have addressed this problem from the perspective of subset selection [7] or finding interesting nodes in a graph [8]. Schaffhausen et. al. [9] propose usage of Semantic Textual Similarity (STS) algorithms to rate the semantic similarity of text sentences submitted by users of a custom open innovation platform. They used human generated scores as benchmark for similarity comparison. Design is unique compared to these in that the ideas can be unstructured, coming from a wide variety of sources and the resultant recommendations can act as seeds to more creative ideas. For design problems, creativity often in-

*Address all correspondence to this author.

volves generating a variety of potential solutions. As one might expect, the goal of any crowd-sourced ideation technique should be to avoid premature convergence on a very limited set of ideas by helping contributors explore the design space

In this paper, we address this problem by proposing a methodology to discover a good set of ideas from a collection of ideas. We define a good set as a small subset of the original ideas that has high quality content, represents the corpus well, and covers diverse topics. It should capture significant themes in the original ideas more efficiently than any random sample can. We view this problem through the lens of recommender systems: 1) we first find a common topic representation of ideas; then 2) we define properties over this representation that encode diversity, quality, and representativeness; and then 3) we optimize over those quantities, providing designers a trade-off between them to find inspiration. This approach is different from related work as it represents the ideas by easily comprehended topics and provides a trade-off front to the designers with many sets of ideas, giving them flexibility for decision making.

To obtain the subset, we first represent the ideas using topic models. Several latent topic discovery techniques such as Latent Dirichlet allocation (LDA) [10] reveal hidden topics for each idea in a document collection. With latent topics uncovered, we derive similarity between different ideas, which we then use to define diversity and representativeness. Lastly, we optimize these measures to find the ideal set.

With the assumption that existing ideas often act as stimuli for new ideas, we apply our technique to OpenIDEO challenges to find recommendations for participants. OpenIDEO is a successful online open innovation community centered around designing products, services, and experiences that promote social impact by building of ideas from distributed individuals [11]. In general, each challenge has a problem description and various stages like: 'Research, Ideas, Applause, Refinement, Evaluation, and Winners.' Our focus in this study is at the 'Ideas' stage, where participants generate and view potential design ideas. In this stage, new participants can review hundreds to thousands of previous ideas to gain inspiration when developing their own ideas; in practice, the number of submissions make exhaustive review (even of the titles) impossible—for a single, medium-sized challenge ($\approx$ 500 ideas) it would take a person over 40 hours to read all idea entries. Because of this, participants often filter by date, the total number of comments, or just pick ideas randomly. Once inspired, participants in a challenge submit new ideas containing text and images, linking to existing ideas that inspired them. Over time, submitted ideas accrue views, applause, and comments as other participants provide feedback [12].

The rest of the paper is organized as follows. The next section presents our proposed approach. The results section justifies our proposed approach through experiments on data from an OpenIDEO challenge. The later sections address research limitations and future work, highlighting the design implications of the paper.

METHODOLOGY

Our approach to recommending a good set of ideas consists of three steps: idea representation, quantifying what makes a "good" set of ideas, and then optimizing the set of ideas. First, we represent each idea as a vector of words and find hidden topics in them. Second, we cluster ideas that have similar topics and use these clusters to define measures of diversity, quality, and representativeness for a good set of ideas. Lastly, an optimization method finds sets of ideas that trade-off diversity, quality, and representativeness.

Representing Ideas

The first step is to computationally represent an idea. Research on representing text documents largely uses the vector space model where a document is expressed by a vector of keyword weights using the bag-of-words model. Those weights are usually calculated using the TF-IDF method [13]. However, the dimensionality of the vector for the TF-IDF method is the number of unique words in the collection after pre-processing, and can be very large even with a moderate scale corpus. To combat this problem, researchers developed various dimension reduction techniques including probabilistic Latent Semantic Analysis (pLSA) [14] and topic modeling [15]. These techniques resolve the curse of dimensionality problem by capturing hidden semantic structure in a document.

Topic modeling—exemplified by Latent Dirichlet Allocation (LDA) [10]—represents a probabilistic model semantic structure in text. In LDA, each document is described as a random mixture over a set of hidden topics where each topic is a discrete distribution over a text vocabulary. Several fields have successfully used LDA for multi-document summarization [16], information retrieval [17], tag recommendation [18] and topic identification [19]. Our approach uses LDA to capture the topic distribution of a design idea. Specifically, we use the topic proportion vector (θ) to represent each idea for a given number of topics.

Defining a "Good" Set of Ideas

Having represented all the ideas in a challenge (V) as a set of topic vectors, we now want to select a subset $S \subseteq V$ that is diverse, high-quality, and representative of V. Before we can define those three metrics, we first need to measure the similarity between ideas. We do this by computing the cosine-similarity between the topic vectors: 1) the estimated θ from LDA denotes the latent topic distribution of each idea; and then 2) since each idea has its own topic distribution vector (θ_i), we apply a cosine similarity measure between any two ideas vectors. Specifically, we compute the similarity between idea i and idea j—$Sim_{i,j}$—by

comparing their topic vectors (θ_i and θ_j):

$$Sim_{i,j} = \frac{\sum_{w=1}^{T} \theta_{w,i} \times \theta_{w,j}}{\sqrt{\sum_{w=1}^{T} \theta_{w,i}^2} \times \sqrt{\sum_{w=1}^{T} \theta_{w,j}^2}} \quad (1)$$

Where T is total the number of topics and $\theta_{w,i}$ is the topic proportion of idea i for topic w. This essentially becomes one if a pair of ideas talk about similar topics or zero if they differ from each other. (Unlike LSA, an idea cannot have a "negative topic" proportion that could cause "negative similarity.")

Diversity Given a way to measure the similarity of ideas to one another, past literature has defined various diversity metrics including Maximum Marginal Relevance (MMR) [20], absorbing random walks [21], and subtopic retrieval [22]. Many of these existing measures are instances of *submodular functions* [23, 24]. Submodular functions naturally model notions of coverage, representation, and diversity [25] and achieve the best results to date on common automatic document summarization benchmarks (*e.g.*, at the Document Understanding Conference [23,24]). We use the diversity reward function proposed by Lin *et al.* [23] for multi-document summarization, which rewards diversity:

$$F_1(S) = -\sum_{k=1}^{K} \sqrt{\sum_{j \in S \cap P_k} \frac{1}{N \times M} \sum_{i \in V} Sim_{i,j}} \quad (2)$$

Here, $V = v_1, ..., v_n$ is the set of all N ideas in a challenge. Subset $S \subseteq V = s_1, ..., s_m$ is the selected M ideas for recommendation given K clusters. P_i, i = 1,...,K is a partition of the ground set V into separate clusters (*i.e.*, $\cup_i P_i = V$ and the P_is are disjoint). This is, an idea can only belong to one cluster. The negative sign converts the objective into a minimization problem for optimization. In Eq. 2, the value $\sum_{i \in V} w_{i,j}$ basically states that the more similar to the corpus an idea is, the more reward there will be by adding this idea to an empty summary set. The square root function makes sure that additional elements from the same cluster have diminishing gains. Hence it automatically promotes diversity by rewarding ideas from clusters which have not yet contributed ideas. This function is monotone non-decreasing and submodular, which means that a scalable greedy optimization scheme has a constant factor guarantee of optimality. We later use this property to substantially accelerate optimization.

To obtain the clusters for the diversity measure in Eq. 2, many clustering algorithms have been proposed, such as K-means, Spectral Clustering and affinity propagation (AP). Among the above algorithms, K-means [26] and Spectral Clustering [27] needs to specify the number of clusters in advance, while Affinity Propagation (AP) need not. We chose Spectral Clustering for our results, however, we later show that the choice of clustering algorithm does not substantially affect our results.

Quality The recommended set of ideas should not only be diverse, but also of high-quality. For any given idea, OpenIDEO has multiple metrics that indicate quality: 1) Applause—users can endorse an idea by pressing the 'Applaud' button; 2) Citation count—users can cite ideas that inspired them, similarly to academic papers; and 3) Comment or View count—each idea tracks the number of comments or views it receives. We use average applause as our measure of quality since OpenIDEO uses applause as their own quality measure during one of their selection stages:

$$F_2(S) = -\sum_{j=1}^{M} \frac{a_j}{M} \quad (3)$$

Here, a_j is the total applauds received by idea i and M is the number of recommended ideas. Applauds is similar to Facebook 'Like' feature, where community members endorse an idea. We did not combine applauds with views and comment count metrics as there is no straightforward way to determine optimum weights for combining these three metrics. For example, it is difficult to argue if receiving more comments is more important as receiving more views. We found that Applauds had a Pearson's linear correlation of 0.65 with views and 0.69 with comment count, so choosing a different quality measure would not substantially alter our results.

Representativeness Given equally diverse and high quality sets of ideas, we would prefer a set that is representative—that captures the central themes of other ideas in the entire collection. We find the central topic—the mean topic vector of all the ideas in the corpus—and then rank ideas by their similarity to this center. For example, if 90% of ideas talk about mobile applications, a representative set of ideas should include ideas closer to that topic, possibly at the expense of other topics. The representativeness of an idea relates to its similarity *to all* other ideas, while the diversity of a set relates to how diverse ideas are compared to each other *within the selected set*.[1]

One way of ranking similarity is TextRank [28] which determines central ideas in the same way that PageRank selects important web pages: ideas 'recommend' similar ideas to the reader. If one idea is similar to many others, it will represent those ideas well. This idea's representativeness, however, also stems from the representativeness of the ideas 'recommending' it. Thus, to get ranked highly and placed in the selected set, an idea must be similar to many ideas that are in turn also similar to many other ideas. We average the representativeness of each idea (p_j) in a set of M ideas to measure the set's representativeness:

$$F_3(S) = -\sum_{j=1}^{M} \frac{p_j}{M} \quad (4)$$

[1] Notice that representativeness and diversity are at odds with one another—we return to this point later in the paper.

Alternative metrics for representativeness include the average cosine similarity of ideas with entire corpus.

Optimizing the Set of Ideas

An ideal set of ideas should balance diversity, quality, and representativeness. We could maximize any one of these three objectives directly by finding the best combination of ideas to recommend subject to a given metric. For all three, however, we need to optimize across multiple, conflicting objectives. This involves finding sets of solutions that represent optimal trade-offs between diversity, quality, and representativeness. We can then use those trade-offs to help designers explore and filter possible ideas.

In practice, you can use any multi-objective optimizer to explore those trade-offs. We chose to use Multi-Objective Evolutionary Algorithms (MOEAs), specifically the MATLAB implementation of the NSGA-II algorithm [29] with binary variables. We generate the initial population randomly with a binary string of length N with only M number of ideas allowed in any string. The binary value indicates whether an idea is in the set or not. The optimizer selects the next generation of the population using a solution's non-dominated rank and distance to the current generation to avoid crowding. Specifically, we use a controlled elitist genetic algorithm [29] with binary tournament selection, uniform mutation, and crossover.

To compare different trade-off fronts, we use the hypervolume measure proposed by Zitzler and Thiele, [30]—defined as the size of the space covered by a trade-off front. Hypervolume increases when one trade-off front is better than another trade-off front. It is frequently used to compare the results of Multi-Objective Evolutionary Algorithms (MOEAs).

RESULTS AND DISCUSSION

To demonstrate our method's effectiveness on a concrete example, we randomly chose a recent challenge from OpenIDEO entitled "How might we make low-income urban areas safer and more empowering for women and girls?" sponsored by The Amplify Program [31].[2] Our aim is to recommend ten diverse, high-quality, representative ideas from a set of 573 submissions it received during its "Idea" stage.

The text content of ideas was extracted using a web-scraping script and stored as text documents. The ideas were pre-processed by boiler-plate removal, segmenting the text, lemmatizing it, and then stemming words using the Porter Stemmer. We ignored all words with inter-document frequency less than 1% and greater than 90%. Pre-processing reduced the vocabulary from 19,628 to 2,977 unique words with a total of 237,862 words in the corpus. We set the LDA hyper-parameters α—the topic distribution smoothness—and η—the topic-word prior—to values recommended in prior literature [32]: $\alpha = 50/T$ and $\eta = 0.1$. Cosine similarity between every pair of ideas was calculated using Eq. 1. We define the number of clusters to be same as number of recommendations, since this makes it easy to verify the maximum diversity solution: it will have one idea in eachof the different clusters. The square root function for each cluster means that to maximize diversity, multiple ideas in same clusters are not desirable. Choosing fewer clusters will lead to a lower objective function value for the same set of ideas as at least two ideas will be within the square root function.

For this case study, we specified ten topics for the topic model. We visualize the topics by their seven highest probability words for each topic:

1. business training work children group social support
2. violence men rights gender sexual community abuse
3. girls schools education program young training self
4. community create working change issue building stories
5. crime using test attacks organisation safety prototype
6. community developing urban area local project governance
7. people area organisation help ways low feel
8. community information using map phones providers help
9. product project using community income water food
10. safety public cities safe spaces bus transportation

These topics relate to women's safety in different ways: Topic 5 talks about violence against women, Topic 3 discusses education, *etc.* We take a sample idea titled "Community Scorecards for Women's Safety." It talks about a score card toolkit that enables community members to prioritize actions and solutions to address women and girls safety issues. This idea is dominated by Topics 5 and 7 in its topic representation. The top words of these two topics include 'crime' and 'community'—just as one would hope from a meaningful topic representation. With this topic representation, we calculate the cosine similarity between ideas and cluster the ideas into ten group using Spectral Clustering for Eq. 2. With the idea topics and clusters, we first separately optimize for the highest quality or most diverse ten ideas, and then we compare those to jointly optimizing both as well as all three metrics.

Maximizing Quality

To obtain a list of ten recommendations, one naive approach would be to recommend the highest quality ideas—the ones with most applause. The highest applauded solutions can be obtained by a simply sorting all ideas and taking the ten ideas with the highest applause. The titles of the obtained solution are:

1. Hack to the (safer) future!
2. From Open Defecation to Improved Sanitation
3. For Women, by Women Taxis
4. Community Scorecards for Women's Safety

[2] We have tested our approach across multiple challenges, achieving similar results, however for ease of exposition we had to pick one to describe in detail for the paper.

5. Menstruation Matters: UPower
6. EyesOnPublicSpaces
7. Leadership Living Center (LLC): Investing in safe spaces to transform communities
8. Voice - An International Media Project for Women
9. RatemyBus: A crowdsourced mobile safety-rating and review app for bus routes and stops
10. Mama Shwari- Cradle of Women Violence Prevention

Although these are the most applauded ideas, some of them are remarkably similar. Idea 3 talks about taxi companies employing low-income women as drivers for women in India. Idea 6 talks about a safety report card for cities which was inspired by their "Board the Bus" campaign in Delhi, India. To avoid harassment, Idea 9 discussed an mobile rating application designed to identify bus stops and bus routes where sexual harassment has occurred. Ideas 3, 6, and 9 are all essentially about reducing transit-related sexual harassment. These ideas share concepts of safety for women around public places and have some common themes in the description. While the ideas themselves have high quality, the set of ideas lacks diversity. Our LDA representation supports this observation: they all have Topic 10 as the highest topic in different proportions and Ideas 6 and 9 have Topic 7 as the second highest.

Maximizing Diversity

Instead of maximizing quality, another naive approach would be to maximize the diversity of the ten ideas. Such a set should have a wide range of solutions to the challenge. Since our diversity function is sub-modular, we can maximize diversity through greedy optimization, resulting in the following titles:

1. Luminescent Ink
2. Empowering a potential victim
3. Grab the big picture
4. Rent a Trained Escort Dog
5. We are more!
6. What about making virtual spaces safer?
7. Women's buddy system built on top of existing charitable community groups
8. Hey, over here!
9. Free comic books featuring fictional, locally-originated heroines.
10. Clear Wayfinding

These ideas cover a wide range of concepts from painting the buildings with luminescent ink so they glow in the dark, campaigning for better media portrayal of Brazilian women, to utilizing trained dogs to escort women. It is possible that this diverse range of topics inspires novel ways of approaching the problem for a new designer looking for inspiration. However, some of these ideas had minimal text and lacked structure and details. Despite being diverse, they are generally sparsely written, and possess zero citations and few applauds or comments. Hence, we do not recommend presenting only the most diverse ideas without any quality check.

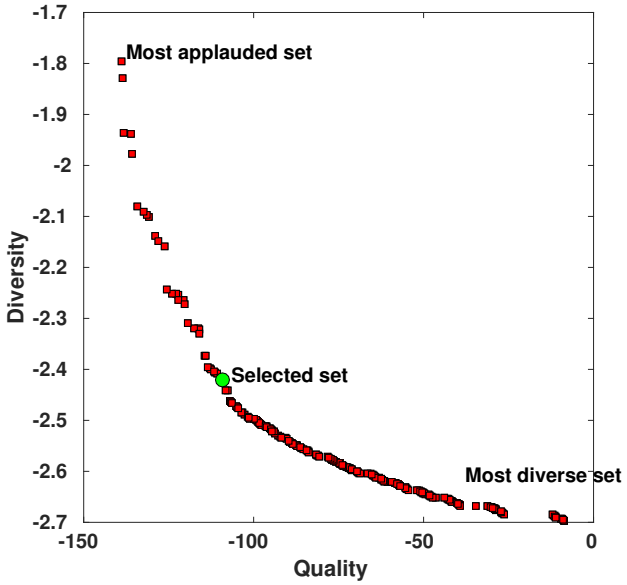

FIGURE 1. Trade-off front between quality and diversity of ideas. The selected solution trades-off 33% diversity for large gain in average applauds

Trading Off Diversity and Quality

In practice, we have to trade-off diversity and applause, hopefully finding a set of high-quality ideas that also maintains diversity. We do this by forming a trade-off front—also called a *Pareto front*—where we can transition from the highest quality ideas to the most diverse ideas, finding a mix of the two. Specifically, we create the front by running NSGA-II using 573 binary variables using a population size of 1000 for 1000 generations. To improve the GA convergence rate, we initialize the population randomly with subset size ten for all genes (total number of 1's in the vector). The single objective solution obtained for diversity and applauds is also introduced in the initial population to further improve the convergence rate.

The GA obtains the trade-off front shown in Fig. 1 with 211 solution sets. Each point on this trade-off front represents a different set of ten ideas found by the optimization algorithm with a different trade-off between diversity and quality. The extremes of the trade-off front represent the highest applause and highest diversity solutions obtained before. One would naturally be interested in understanding which ten documents get selected for every point on the front. Fig. 2 shows the actual ideas (represented

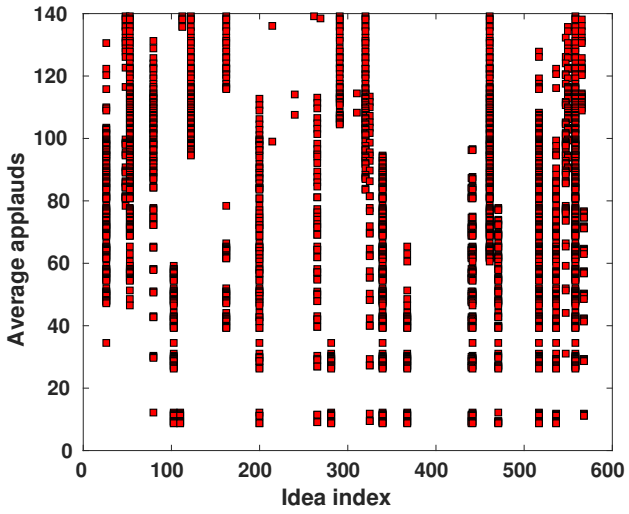

FIGURE 2. Ideas selected in different solution sets on the trade-off front between applauds and diversity. The figure shows only a small set of 33 unique lines appear on trade-off front. On the top are ideas selected for high applauds in the trade-off front, while bottom of the figure has ideas with high diversity

by idea index) in the sets of the various points on the pareto front plotted against the average applauds—as you move from the top of Fig. 2 to the bottom you are moving along the curve in Fig. 1 from the upper-left to the bottom-right. The documents selected previously by our quality maximization are at the top row of this figure while diversity maximization ideas discussed before are at the bottom most row. An interesting insight from Fig. 2 is that there is only a small set of 33 ideas which occur in different combinations to generate the entire trade-off front. We call this set the *trade-off set*. This shows that out of original 573 ideas, only 33 ideas provide the full range of recommendations from high quality to high diversity.

Normally, a designer would prefer recommendations from somewhere in the center of the trade-off front to obtain a balance between quality and diversity. We notice in Fig. 1, that moving away from trade-off solutions around 100 applauds leads to large loss in one objective for small gains in the other objective. Hence for this particular case a solution around 100 applauds might be preferable. For the purposes of demonstration, we looked at the total comments received by each set as a decision making criteria to select a final set with 109.1 average applauds. However, the user can use any external decision making criteria to select a solution on the trade-off front—as we saw in Fig. 2, the changes are not substantial.

The titles for the selected solution are as follows:

1. The Laundry Lab: creating new possibilities one load of laundry at a time
2. From Open Defecation to Improved Sanitation
3. KUPRI: Patching communities.
4. Community Scorecards for Women's Safety
5. Grab the big picture
6. Voice - An International Media Project for Women
7. RatemyBus: A crowdsourced mobile safety-rating and review app for bus routes and stops
8. Community Concierge Program
9. Red Chili Powder Filled Glass Bangle for Women's Self-defense
10. Mama Shwari - Cradle of Women Violence Prevention

The above recommended solution set has ideas talking about different topics—as measured by topic clusters—which is a considerable improvement over the most applauded solutions. Likewise, the average quality of ideas is an order of magnitude higher than the most diverse solution—as measured by applause. This balances quality and diversity. The ideas obtained above discuss a range of concepts from women's sanitation, to mobile applications for bus safety, to physical bangles filled with chili powder as a defense mechanism, among others (see Table 1 on Page 10).

Trading Off Diversity, Quality, and Representativeness

Having obtained a trade-off between diversity and quality for ideas, one may argue that these ideas may not be representative. As hypothetical example, assume 500 of the 573 ideas were based on developing mobile apps to support women safety while 10 ideas were on sanitation, and that most users applauded sanitation ideas. In such a scenario, our most applauded solutions will be sanitation ideas, while the most diverse set of ideas will have one idea on mobile applications and other ideas from the remaining topics. Such recommendations will not be representative of the entire corpus, which is clearly biased towards mobile apps in this case. If we could maintain similar diversity and quality, we would prefer sets of ideas which represented the corpus well—in this case encouraging more app ideas. To do this, we run three-objective optimization to obtain the trade-off surface.

Figure 3 shows a two-dimensional projection of the trade-off between the three objectives; diversity is shown with a color scale. This result was counter-intuitive: for a given quality on trade off front, minor gains in representativeness lead to a large loss in diversity. In other words, quality and representativeness linearly vary. Highly applauded ideas were unique, hence less representative, while ideas similar to the entire corpus did not receive much applause.

This result could be interpreted as users preferring to applaud unique ideas on OpenIDEO. Another possible interpretation could be that similar good quality ideas dilute or share applause among one another. In our case, including representativeness does not add much value to our final recommendations; setting a quality requirement essentially sets representativeness.

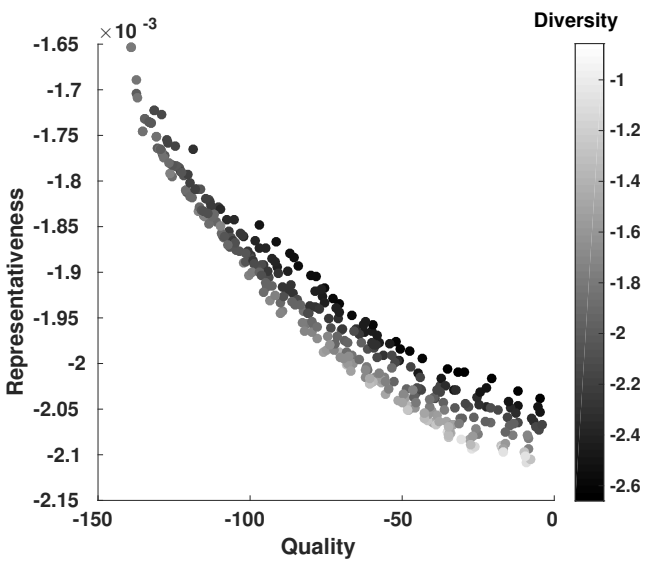

FIGURE 3. Trade-off between applauds and TextRank. For a given quality, small improvements in TextRank result in large loss of diversity.

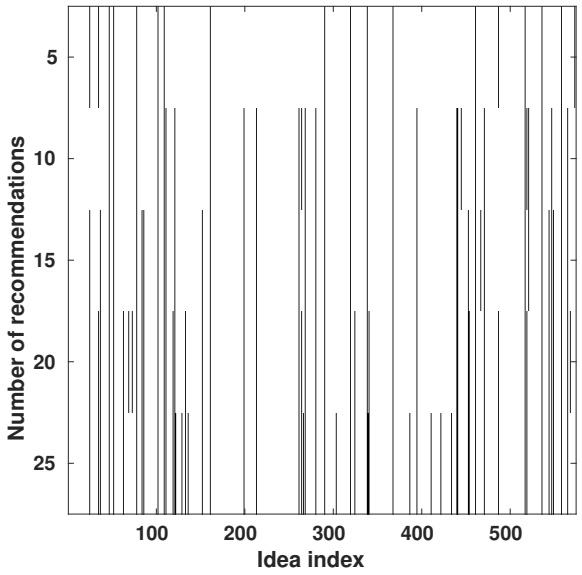

FIGURE 4. Trade-off set showing ideas selected for different number of recommendations. Vertical lines from top to bottom show documents always selected irrespective of number of recommendations.

The Effect of Model Parameters

In the previous cases, we provided ten recommendations for diversity and quality maximization, then showed how to trade off two and three objectives. We chose ten topics and used Spectral Clustering with ten clusters in all results so far. In this section, we analyze the impact of those choices.

Different Number of Recommendations We tested our methodology for 5, 10, 15, 20 and 25 recommendations for ten topics and ten clusters, finding the trade-off set for all the five cases. For single objective optimization using a greedy algorithm, one would expect that moving from 10 to 11 recommendations would include the previous 10 for both quality and diversity. These algorithms work by sequentially adding one element to the set which maximizes the total reward value.

Interestingly, we observed similar behavior for trade-off fronts obtained by global optimization. The trade-off front obtained for N recommendations largely overlapped with all M recommendations, where $M > N$. This means that one can optimize for larger number of recommendations and if a user requires fewer recommendations, the new set will be a subset of larger set. This property can also be used to accelerate GA convergence by initializing the population with larger set members. The idea indices in the trade-off set for the five idea case is shown in Fig. 4.

Clustering algorithm Previously, we had chosen Spectral Clustering for our analysis. In this section, we test four different clustering algorithms for calculating diversity: K-Means, Spectral Clustering, Affinity Propagation (AP), and Highest Topic Allocation. The latter method is hard assignment of an idea to its highest topic, implying that the topic the idea talks about the most defines the cluster it belongs to. Cosine similarity is used as a distance measure for Spectral Clustering and AP. For first three methods, the number of clusters are defined as ten while AP automatically finds the number of clusters.

Figure 5 shows the trade-off fronts obtained for the four cases. It can be seen that AP obtains slightly better diversity for same number of applauds on the trade-off front compared to other three methods, while K-Means in general does not perform well. This can be explained by the observation, that AP found 33 clusters in the corpus, while we chose ten clusters for the other three methods. This leads to higher probability of ideas lying in unique clusters hence improving the diversity objective function values for AP. For the purposes of our experiments, we chose the Spectral Clustering method, as AP created a large number of clusters, artificially improving the diversity, by creating clusters between ideas that did not appreciably differ. For example, AP considers all the highest applause solutions in different clusters despite the fact that, as we discussed previously, the ideas share common themes and lack diversity. We found that using cosine similarity as a clustering distance metric improved trade-off fronts compared to euclidean distances.

In general, there were two important takeaways from analyz-

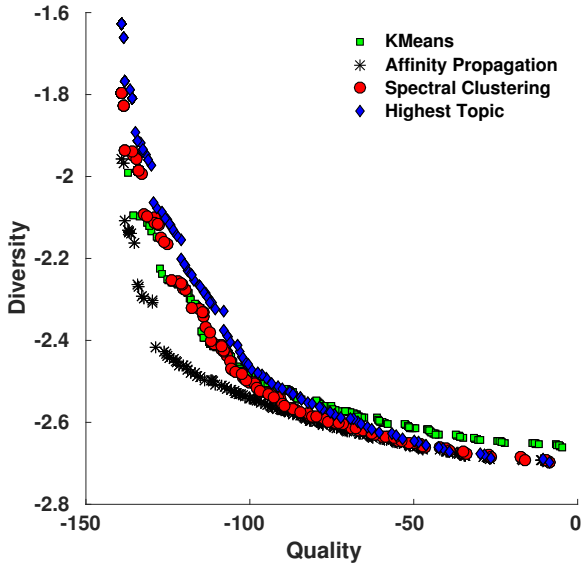

FIGURE 5. Trade-off front between applause and diversity for different clustering methods. AP performs better for high applauds as it has larger number of clusters.

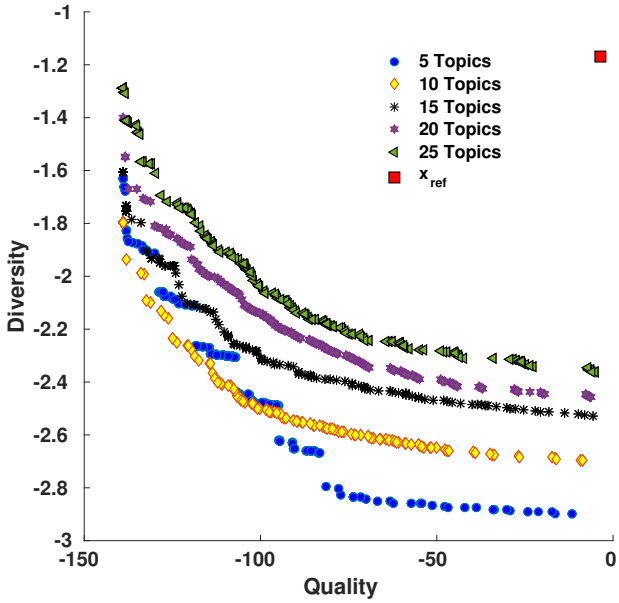

FIGURE 6. Trade-off front for different number of topics for quality and diversity measures. Moving to lesser topics generally improve diversity

ing clustering algorithms. The choice of clustering method only created minor changes in the final trade-off front as seen in Fig. 5. Secondly, the most diverse solution found by different methods shared many common ideas. Hence, different algorithms are able to distinguish between ideas based on cosine similarities.

Number of Topics We analyze the impact of the number of topics on the trade-off front. Intuitively, the number of topics define the dimensional reduction of our ideas. This affects the cosine similarity between ideas which is essential for correct assessment of diversity and pagerank. If the similarity matrix is incorrectly estimated, then the clustering algorithm may allocate essentially different ideas to the same cluster or vice versa. This can occur if we chose too few or too many topics.

To estimate the number of topics, we propose a hypervolume based approach. The basic assumption is that solutions on a trade-off front with lower objective values are better. Fig. 6 shows the trade-off fronts for five cases with number of topics 5, 10, 15, 20 and 25. In general, increasing the number of topics leads to a trade-off front which has lower diversity for same number of applauds. Hence we estimate the number of topics, by calculating the trade-off front and choosing the one which dominates the other fronts. This is the same as maximizing the hypervolume of the trade-off front calculated with respect to a reference point x_{ref}. The mean hypervolume from ten runs is shown in Fig. 7 as there can be variations caused by random initialization of LDA. The hypervolume initially increases with number of topics and then goes down as the number of dimensions goes up. The ideal number of topics in this case will be around six, which obtains the highest hypervolume. Other approaches like LDA perplexity calculation [33] or using Hierarchical Dirichlet Process [34] can also be used to estimate the number of topics.

Applicability to Different Domains

For our collection of 573 ideas, an average reader (at 200 words per minute) would need 40 hours to actually read all ideas, causing significant fatigue and loss of attention. Applying our method for ten recommendations reduces this reading time to 40 minutes while still capturing good quality and diverse ideas. This shows the benefit of applying such a method to any collections of design ideas.

The domain of designs considered in our experiments are those from OpenIDEO, which often produces diverse solutions to problems that span products, services, policy interventions. However, our method equally applies to any collection of design ideas expressed as text. In its current form, it cannot be applied to sketches and images. However, one can note that the primary usage of text was to find topics, which in turn calculated similarity between ideas. Hence, any other method which provides similarity between ideas based on text or image data can be directly plugged into our method. An interesting area of further study can be to use automatic image annotation [35] to extract

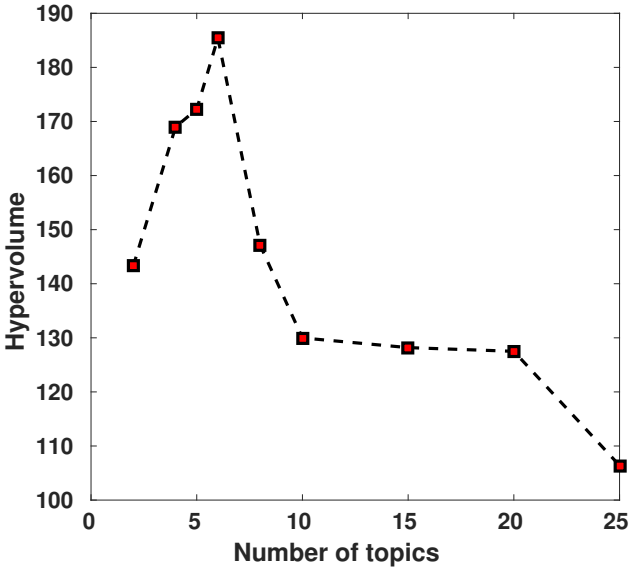

FIGURE 7. Average hypervolume for trade-off front for different number of topics. It shows that very few or large numberof topics are an inadequate representation

text keywords from images and add them to the text content of the idea before applying topic modeling.

In its current form, designers can directly apply the method to their collection of text ideas and reduce the number of generated designs. If the design ideas do not have available quality metric like applause, they can use representativeness and diversity to obtain the trade-off front.

CONCLUSION

In this paper, we proposed a method to recommend a small subset of ideas from a large corpus of ideas. Specifically, we demonstrated our method by optimizing the quality, diversity, and representativeness of 10 recommended ideas out of 573 possible ideas from an OpenIDEO challenge. We conducted a parametric analysis across the number of recommendations, the choice of clustering algorithm, and the number of topics—changes to these parameters did not fundamentally limit our method's ability to recommend diverse, high quality ideas.

Future research avenues include: 1) better ways of representing networks of design ideas, such as through Relational Topic Models [36]; 2) improving quality metrics so that they leverage both human assessment (*e.g.*, applause) and text content; and 3) recommending ideas based not only on content, but on a designer's expertise or preferences.

Our findings have several implications both for recommending ideas and studying ideation at large scale. First, Fig. 2 showed that, out of 573 ideas, only 33 unique solutions appeared across any portion of the pareto front, from high-quality to high-diversity. This implies that, even without picking a location on the pareto front, we can achieve substantial compression in the "minimal set" of inspiring ideas a designer might consider—roughly 6% in our example. Second, when trading off diversity and quality, we found that maximizing diversity without considering quality produced less useful ideas than considering the combination. This implies that we need better automated quality metrics for ideas—similar to those researchers have proposed for diversity or variety—if we hope to scale up our ability to evaluate or inspire creative ideas. Third, our diversity and representativeness measures rely on a topic-based representation of ideas. This was a useful—though crude—way to summarize a design idea. While a design clearly has a deeper structure to it than a technique like pLSA or LDA could hope capture, our results demonstrate the practical usefulness of topic representations for categorizing large collections of ideas. Combining those techniques with more structured design formalisms could improve our ability to accurately recommend ideas.

Ultimately, strengthening our ability to understand large design collections not only could improve how we design new products, but could provide deeper insights into how, why, and what we design.

REFERENCES

[1] Pauling, L., and Kamb, B., 2001. *Linus Pauling: selected scientific papers*, Vol. 2. World Scientific.

[2] Shah, J. J., Kulkarni, S. V., and Vargas-Hernandez, N., 2000. "Evaluation of idea generation methods for conceptual design: effectiveness metrics and design of experiments". *Journal of Mechanical Design,* ***122***(4), pp. 377–384.

[3] Kudrowitz, B. M., and Wallace, D., 2013. "Assessing the quality of ideas from prolific, early-stage product ideation". *Journal of Engineering Design,* ***24***(2), pp. 120–139.

[4] Green, M., Seepersad, C. C., and Hölttä-Otto, K., 2014. "Crowd-sourcing the evaluation of creativity in conceptual design: A pilot study". In ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, pp. V007T07A016–V007T07A016.

[5] Von Hippel, E., 2005. "Democratizing innovation: The evolving phenomenon of user innovation". *Journal für Betriebswirtschaft,* ***55***(1), pp. 63–78.

[6] Chiu, I., and Shu, L., 2012. "Investigating effects of oppositely related semantic stimuli on design concept creativity". *Journal of Engineering Design,* ***23***(4), pp. 271–296.

[7] Pan, F., Wang, W., Tung, A. K., and Yang, J., 2005. "Find-

TABLE 1. OpenIDEO ideas on trade-off front

Title	Description
The Laundry Lab creating new possibilities one load of laundry at a time	Improving english skills of immigrants during laundry time
From Open Defecation to Improved Sanitation	Building appropriate household sanitation facilities to solve women's safety and healthy
KUPRI: Patching communities.	Encourage communities collaboration through the development of a value-collecting backpack line that brings economical safety
Community Scorecards for Women's Safety	To identify safety issues using a scorecard system and prioritize action
Grab the big picture	Communications around encouraging strong moral values in business
Voice - An International Media Project for Women	An international media project that creates a direct channel between women in need and a specific, targeted audience
RatemyBus: A crowdsourced mobile safety-rating and review app for bus routes and stops	Mobile rating application designed to identify bus stops and bus routes where sexual harassment has occurred
Community Concierge Program	Empowering women by providing them access to basic training and supporting them in becoming financially independent
Mama Shwari- Cradle of Women Violence Prevention	It is about women adopting communities strategies through family parenting sessions, support groups
Red Chilli Powder Filled Glass Bangle for Women's Self-defense	Bangles filled with chilli as safety device

ing representative set from massive data". In Data Mining, Fifth IEEE International Conference on, IEEE, pp. 8–pp.

[8] Langohr, L., et al., 2014. "Methods for finding interesting nodes in weighted graphs". PhD thesis, University of Helsinki, 6.

[9] Schaffhausen, C. R., and Kowalewski, T. M., 2015. "Large-scale needfinding: Methods of increasing user-generated needs from large populations". *Journal of Mechanical Design,* ***137***(7), p. 071403.

[10] Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003. "Latent dirichlet allocation". *the Journal of Machine Learning Research,* ***3***, pp. 993–1022.

[11] Fuge, M., Tee, K., Agogino, A., and Maton, N., 2014. "Analysis of collaborative design networks: A case study of OpenIDEO". *Journal of Computing and Information Science in Engineering,* ***14***(2), Mar., pp. 021009+.

[12] Fuge, M., and Agogino, A., 2014. "How online design communities evolve over time: the birth and growth of OpenIDEO". In ASME International Design Engineering Technical Conferences, ASME.

[13] Salton, G., and Buckley, C., 1988. "Term-weighting approaches in automatic text retrieval". *Information Processing & Management,* ***24***(5), pp. 513–523.

[14] Hofmann, T., 1999. "Probabilistic latent semantic indexing". In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp. 50–57.

[15] Blei, D. M., and Lafferty, J. D., 2009. "Topic models". *Text Mining: Classification, Clustering, and Applications,* ***10***(71), p. 34.

[16] Haghighi, A., and Vanderwende, L., 2009. "Exploring content models for multi-document summarization". In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp. 362–370.

[17] Wei, X., and Croft, W. B., 2006. "LDA-based document models for ad-hoc retrieval". In Proceedings of the 29th Annual International ACM SIGIR conference on Research and development in information retrieval, ACM, pp. 178–185.

[18] Krestel, R., Fankhauser, P., and Nejdl, W., 2009. "Latent dirichlet allocation for tag recommendation". In Proceedings of the third ACM conference on Recommender systems, ACM, pp. 61–68.

[19] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P., 2004. "The author-topic model for authors and documents". In Proceedings of the 20th conference on Uncertainty in Artificial Intelligence, AUAI Press, pp. 487–494.

[20] Carbonell, J., and Goldstein, J., 1998. "The use of MMR, diversity-based reranking for reordering documents and producing summaries". In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp. 335–336.

[21] Zhu, X., Goldberg, A. B., Van Gael, J., and Andrzejewski, D., 2007. "Improving diversity in ranking using absorbing random walks.". In HLT-NAACL, Citeseer, pp. 97–104.

[22] Zhai, C. X., Cohen, W. W., and Lafferty, J., 2003. "Beyond independent relevance: methods and evaluation metrics for subtopic retrieval". In Proceedings of the 26th Annual In-

ternational ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM, pp. 10–17.

[23] Lin, H., and Bilmes, J., 2011. "A class of submodular functions for document summarization". In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, pp. 510–520.

[24] Lin, H., and Bilmes, J. A., 2012. "Learning mixtures of submodular shells with application to document summarization". *arXiv preprint arXiv:1210.4871*.

[25] Fuge, M., Stroud, J., and Agogino, A., 2013. "Automatically inferring metrics for design creativity". *ASME Paper No. DETC2013-12620*.

[26] Manning, C. D., and Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*, Vol. 999. MIT Press.

[27] Ng, A. Y., Jordan, M. I., Weiss, Y., et al., 2002. "On spectral clustering: Analysis and an algorithm". *Advances in Neural Information Processing Systems, **2***, pp. 849–856.

[28] Mihalcea, R., and Tarau, P., 2004. "Textrank: Bringing order into texts". Association for Computational Linguistics.

[29] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T., 2002. "A fast and elitist multiobjective genetic algorithm: Nsga-ii". *Evolutionary Computation, IEEE Transactions on, **6***(2), pp. 182–197.

[30] Zitzler, E., and Thiele, L., 1999. "Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach". *IEEE Transactions on Evolutionary Computation, **3***(4), pp. 257–271.

[31] OpenIDEO - how might we make low-income urban areas safer and more empowering for women and girls? - ideas. `https://challenges.openideo.com/challenge/womens-safety/ideas`. (Visited on 01/15/2016).

[32] Griffiths, T. L., and Steyvers, M., 2004. "Finding scientific topics". *Proceedings of the National Academy of Sciences, **101***(suppl 1), pp. 5228–5235.

[33] Boyd-Graber, J., Mimno, D., and Newman, D., 2014. "Care and feeding of topic models: Problems, diagnostics, and improvements". *Handbook of Mixed Membership Models and Their Applications; CRC Press: Boca Raton, FL, USA*.

[34] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M., 2012. "Hierarchical dirichlet processes". *Journal of the american statistical association*.

[35] Zhang, D., Islam, M. M., and Lu, G., 2012. "A review on automatic image annotation techniques". *Pattern Recognition, **45***(1), pp. 346–362.

[36] Chang, J., and Blei, D. M., 2009. "Relational topic models for document networks". In International Conference on Artificial Intelligence and Statistics, pp. 81–88.