# Towards Computational Assessment of Idea Novelty

Kai Wang
School of Management and Marketing, Kean University
kaiwan@kean.edu

Boxiang Dong
Department of Computer Science, Montclair State University
dongb@mail.montclair.edu

Junjie Ma
Global Business School, Kean University
maju@kean.edu

## Abstract

*In crowdsourcing ideation websites, companies can easily collect large amount of ideas. Screening through such volume of ideas is very costly and challenging, necessitating automatic approaches. It would be particularly useful to automatically evaluate idea novelty since companies commonly seek novel ideas. Three computational approaches were tested, based on Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and term frequency–inverse document frequency (TF-IDF), respectively. These three approaches were used on three set of ideas and the computed idea novelty was compared with human expert evaluation. TF-IDF based measure correlated better with expert evaluation than the other two measures. However, our results show that these approaches do not match human judgement well enough to replace it.*

## 1. Introduction

Companies are typically under pressure to improve existing offerings and come up with new ones. This task of innovation has traditionally been given to internal development and marketing teams. Recently, however, more and more companies seek to collect ideas from a large number of people through an open call, typically on the Internet, i.e., through crowdsourcing [4,16]. For example, mystarbucksidea.com collects ideas for improving the products and services of Starbucks while the Dell IdeaStorm website serves the same purpose for Dell. Although this approach enables the fast collection of many ideas, the task of evaluating a large number of ideas and selecting the best ones is very challenging. The number of ideas submitted is more than 100, 000 in mystarbucksidea.com [23] and more than 28,000 for Dell IdeaStorm (www.ideastorm.com). It is quite difficult to have some experts evaluate the novelty and quality of thousands of ideas. In addition, despite the intention of searching for novel ideas [30], companies have the tendency to focus on familiar ideas when they are overwhelmed by a large number of ideas [26]. In other words, manually selecting the most innovative ideas from a large pool is less effective and therefore reduces the value of crowdsourcing innovation.

It would be very helpful to automate, or partially automate, the evaluation of creative ideas. However, since ideas are usually in the form of written text, a type of unstructured data, the computational evaluation of such data is not straight forward [17]. There is a lack of studies directly addressing this topic, especially in the domain of management and information systems research. There are some approaches that have been used to measure the semantic distance between ideas: Latent Semantic Analysis (LSA) [14,36] and Latent Dirichlet Allocation (LDA) [7,33]. These techniques are natural language processing techniques that can use word frequencies and related matrices to compute semantic distances among ideas. Since idea novelty is essentially the degree to which an idea is semantically distant from other ideas, these techniques can potentially evaluate idea novelty. Idea or proposal novelty has also been evaluated with keywords [8,9] and term frequency–inverse document frequency (TF-IDF) [28] but such assessments have not been validated through the comparison with human judgment. In summary, there are some existing approaches that can be used in evaluating idea novelty in a more automatic way. However, these methods need to be compared to human judgment and to each other. Our research question therefore is: which computational method tends to match the novelty evaluation of human experts better? The following section will review the literature on idea evaluation in crowdsourcing, with a focus on existing methods that may automate idea novelty evaluation. Then the experiments collecting ideas and the methods to evaluate idea novelty are described, followed with the results and discussion.

## 2. Background

Although outsourcing a task to a crowd is not a new phenomenon, using the Internet for this purpose is a

URI: https://hdl.handle.net/10125/59531
ISBN: 978-0-9981331-2-6

relatively recent development. The Internet makes it much easier to reach out to many people and tap into diverse knowledge and perspectives. Since the diversity in perspectives and knowledge is very useful for innovation, crowdsourcing becomes an important choice for managers seeking innovations [3,34]. For example, many companies have designated websites or online communities for crowds to generate ideas on improving or creating products and services. In this type of crowdsourced ideation, it is common to collect hundreds or thousands of ideas. The evaluation of the large amount of ideas collected is a major challenge for companies. First, it is difficult and costly to have a few experts evaluate many ideas. Expert evaluation is the standard for assessing creative ideas [2,7]. However, since idea evaluation requires complicated cognitive effort, large scale idea assessment can easily lead to fatigue and poor performance [17]. Relatedly, even though companies seek innovative ideas in crowdsourcing, they tend to focus on familiar ideas, instead of novel ideas, due to the inability to attend to large number of ideas (i.e., bounded rationality) [26]. Second, having crowds (such as Amazon Mechanical Turk workers and online communities) evaluate ideas is not necessarily reliable. While crowd evaluation is cost-efficient, it is dependent on whether the crowd has the relevant expertise and motivation to make the most accurate assessment. For example, it was found that crowd members in ideastorm.com commonly underestimated the costs of implementing their ideas and it was slow for them to learn about the company's cost structure [18]. It has also been found that crowds are much better at identifying bad ideas than selecting good ideas [21]. Consequently, in addition to human judgment, it is sensible to look for alternative methods to evaluate large number of ideas collected in crowdsourcing. Creative ideas are commonly evaluated in two dimensions: novelty and usefulness (or appropriateness) [10,15]. Idea novelty and idea usefulness are not equivalent: an idea can be novel and useless, and vice versa. Therefore, idea evaluation contains more than just novelty assessment. However, we want to focus on idea novelty assessment for the following reasons. First, crowdsourcing projects usually seek novel ideas [30], instead of common ideas. In other words, unoriginal ideas are rarely worth the time and effort spent on launching and managing a crowdsourcing project. Second, novelty evaluation and usefulness assessment are distinct and seem to demand different methods. Idea novelty is more about being distinct from existing ideas, and computational methods are useful in making this judgment. The evaluation of idea usefulness, however, is likely to be domain dependent. Domain-generic computational methods seem less promising in assessing idea usefulness. In the long run, it would be ideal to computationally assess both idea novelty and usefulness. For now, it is practical to start with the evaluation of idea novelty first.

The management and information systems literature indeed has explored alternative methods of idea novelty assessment. Some researchers use LSA [22] (or LSI, latent semantic indexing) to estimate semantic distances as a proxy for novelty assessment. LSA is a technique in natural language processing. In the context of processing ideas, LSA can represent all the ideas in a word by idea matrix, where each word takes a row and each idea takes a column. The values in the cells in the matrix are word frequencies. Then singular value decomposition is conducted to obtain a lower-dimension approximation of the original matrix by maintaining only the factors that account for the most variance. LSA only identifies a linear subspace in the space of TF-IDF features that represents the most variance in the documents. Therefore, it cannot deal with the polysemy among all the documents. In the reconstructed word-idea space, the problem of synonymy is minimized. This is because in LSA the meaning of individual words is inferred from the context of occurrence and synonyms tend to appear in the same context [24]. Clustering techniques can then be used on this low-dimensional space to categorize ideas [24] and the categorization can be used to estimate idea novelty [17]. Specifically, previous ideas are put into clusters and a new idea is assigned into an existing cluster based on semantic similarity. A score is calculated as the number of previous ideas in the cluster to which the idea was assigned, divided by the total number of previous ideas [17]. The lower the score, the more novel the idea. It turns out that moderately novel ideas were less likely to be selected by the studied company while highly novel and highly common ideas had a better chance to be implemented [17]. In another study [14], the idea novelty of divergent thinking tasks (alternative use of a common object) was assessed by computing the cosines between the vector representing an idea and the vector representing the definition of the common object. A novel idea would result in a small cosine value, indicating the dissimilarity between the idea and the object definition. LSA distances were correlated with originality ratings well (correlation coefficient around 0.2 and above) in two out of four divergent thinking tasks.

A second method used for idea novelty assessment is LDA. LDA is a topic modeling approach. LDA has been successfully used in analyzing topics in social media [32] and facilitating information retrieval from academic articles [12]. It can be considered as identifying "the hidden structure that likely generated the observed collection" [5]. Each document is viewed as a mixture of latent topics. All topics, in different

proportions, are used in a document. LDA introduces a Dirichlet prior on document-topic distributions and uses Bayesian statistical learning algorithms to infer the topical structure of the corpus from the word co-occurrence patterns [5,7,25]. LDA algorithm can produce two outputs: a list of topics, each with a vector of words associated, and a list of documents, each with a vector of topic distribution [20]. These document-topic and topic-word distributions represent the generative probabilistic process of the documents. It is effective in handling the synonymy and polysemy. [20] used LDA to find highly novel patents by identifying the first patent to have a significant weight on a specific topic. Some researchers have used LDA for related purposes, even though not for novelty assessment itself. [7] used LDA to calculate the conceptual similarity among design ideas by computing the cosine between their topical mixtures (vectors of topic weights). The correlation of the cosine similarity with the human-judged similarities was .54 and .51 for two sub-samples. Using the same LDA approach, [31] calculated the similarity among companies based on the unstructured texts of company descriptions. Companies in the same category (categories are defined by the database) had a mean business proximity value twice as large as companies from different categories.

A third method for novelty assessment is TF-IDF. TF-IDF is a term weighting method used in information retrieval [29]. The number of occurrence of a term in a document is normalized by the document length and multiplied with the inverse document frequency of the term (IDF). Consequently, a term used a lot in one document but rarely used in other documents receives a high TF-IDF value. TF-IDF is effective in finding the important and innovative words. However, it fails to capture any intra- or inter-document statistical structure information. [28] used the sum of TF-IDF values for all terms to estimate the novelty of an idea. Relatedly, logarithmic calculations [26] and cosine similarity [38] based on TF-IDF values are used to assess idea novelty.

There are still other methods of novelty assessment that do not rely on human judgment. However, these methods require either some training set or pre-existing set of keywords or categories. Toubia and Netzer [33] built a semantic network based on a training set (of ideas or Google search results) on a particular idea generation topic where nodes represent words and the weight of an edge is the scaled co-occurrence of the two words. Then a semantic subnetwork was constructed for each idea based on its own set of words. The researchers showed that ideas with semantic subnetworks that have a more prototypical edge weight distribution (i.e. similar to the edge weight distribution of the overall semantic network) are judged as more creative. [19] identified highly novel patents by finding the first combination of two patent subclasses (pre-coded in the patent database). Similarly, [8] and [9] evaluated research proposal novelty by the rarity of the combination of pre-defined keywords. While these novelty assessment methods can be effective, the requirement of pre-existing keywords, categories or training sets makes them quite complicated and often infeasible.

It is important to compare these computational methods with the traditional method of human expert judgment: the standard method of the evaluating creative ideas [2]. However, there is a lack of such studies. While [7] and [14] showed some correlation between human judgment and their computational approaches (LSA and LDA, respectively), there is no comparison across different computational approaches. Consequently, we do not know which approach tends to be better. [38] indeed compared different approaches based on TF and TF-IDF in how closely their results matched expert idea selection. However, they did not test other approaches that are commonly used and more advanced (such as LSA and LDA). In addition, it is worth noting that some studies used idea implementation by the company [17] or expert idea selection [38] as the ground truth of idea innovativeness for the validation or testing of computational methods. However, whether an idea is selected or implemented may not be the ground truth of idea innovativeness or novelty. There are three major reasons. First, in idea selection, companies need to consider many factors other than idea novelty, such as market conditions and cost structure. Second, companies might be inherently risk averse such that they select familiar ideas to implement. Third, it is documented that companies could be overwhelmed by the task of screening through thousands of ideas and such cognitive overload may lead to a preference towards familiar instead of novel ideas [26]. These limitations are minimized if experts only need to give novelty scores to manageable amount of ideas. Therefore, it is worth testing different computational methods of evaluating idea novelty using expert scoring (instead of selection) as the ground truth while keeping the number of ideas manageable so that experts are not overwhelmed.

## 3. Methods

To test the computational methods of idea novelty assessment, we collected many ideas and obtained novelty ratings by experts. We used Amazon Mechanical Turk (www.mturk.com) to employ crowd workers for idea generation for three tasks. The first task was to generate creative ideas about designing a mobile app. The app was to be used by college students as an alarm clock. The second task was to generate creative

ideas about a mobile app helping college students improve physical fitness. The third task was to ideate about a TV advertisement promoting public transportation in order to improve the environment and support sustainability. All the three topics are realistic crowdsourcing topics. Three separate groups of crowd workers worked on the three tasks. Each person generated one idea and earned one US dollar. There are 200 alarm clock app ideas with the average length of 555 characters (SD=287). There are 240 fitness app ideas averaging 586 characters (SD=312). Three hundred TV ad ideas were collected averaging 307 characters (SD=191). Each set of the app ideas was evaluated in novelty by two mobile app developers with at least 5 years of professional experience. Two distinct pairs of experts evaluated the alarm clock app ideas and the fitness app ideas. The TV advertisement ideas were evaluated by two experts with at least 5 years of professional experience with TV advertising. These raters were instructed to look through existing apps or video advertisements before the idea evaluation. Novelty is defined as the degree to which an idea is rare and unique using a Likert scale of 1 to 7 (1 being not novel at all, 7 being highly novel). The experts show reasonable level of agreement in the ratings (Intraclass correlation coefficient, ICC(2,2)=0.69, 0.75, and 0.62, for alarm clock app ideas, fitness app ideas, and TV advertisement ideas, respectively). Therefore the scores from two experts were averaged to obtain the novelty ratings. These expert-judged novelty ratings were used as the ground truth for computational methods to match.

Three computational approaches to novelty assessment were used. The first method is based on TF-IDF. For each term in a document, the term frequency is normalized by the document length and then multiplied by inverse document frequency of the term [28]. Like [28], the TF-IDF values for all terms in an idea were summed as the measure for novelty. The second and third method are based on LSA and LDA respectively. In each method, a vector is calculated to represent each idea. In LSA, it is the vector in the LSA space corresponding to an idea. In LDA, it is the vector describing the topic distribution of an idea. In both methods, the vectors representing ideas were used to calculate a cosine similarity between ideas. Subtracting this similarity value from one resulted in semantic distance between ideas. We calculated the average distance from one idea to all the other ideas as a novelty measure (noted as *average distance*). We argue that if an idea is semantically distant from all other ideas, it is novel. Another approach was to calculate an "average idea vector" by averaging all the vectors representing ideas. Considering this vector as the center of the semantic space, we calculated the distance between each idea and this center by subtracting cosine similarity from one. We used this distance as a second measure of novelty for both the LSA and the LDA methods (noted as *distance to average*). We used the *lsa* package in R for LSA. We used the packages *tm* and *topicmodels* in R for LDA. We trained the LDA model using Gibbs sampling with 2,000 iterations, where 5 random starts are repeated and the first 4,000 iterations are omitted in order to stabilize the model. Stemming and standard English stop words from the packages were used.

Lastly, for all three ideation tasks, we also collected crowd evaluation of idea novelty as another assessment method to be compared. Each idea was evaluated by twenty Mechanical Turk workers on the same scale of idea novelty (1 to 7) as used by experts. It was reported that twenty crowd ratings per idea are enough to result in stable idea ranking [27]. Each worker was paid 3 US cents for every idea evaluated. The agreement among crowd ratings is good (ICC(1,20) is at least 0.70 in all the three tasks). Twenty scores for each idea were averaged to obtain the novelty score. The novelty values obtained by all the methods were compared to the ground truth (expert evaluation) through Pearson correlation. A good method of idea novelty assessment should highly correlate with expert evaluation.

## 4. Results

In using LSA and LDA for novelty assessment, we have two measures: *average distance* and *distance to average*. For both LSA and LDA, for all three ideation tasks, these two measures have correlation coefficients consistently above 0.98. Therefore, these two measures are almost identical in their ability to differentiate ideas in novelty. Thus, we use only one measure: average distance, and ignore the other measure.

In using LDA, an important parameter is the number of topics (k). Previous researchers find that even though best-fit models tend to have a large topic number, having smaller topic numbers can better result in topics of distinct meanings [20]. Considering that we have only 200 to 300 ideas per task, we test three different k values: 10, 20, and 30. The descriptive statistics of idea novelty and the correlations among the ground truth and the different measures are in Table 1. The LDA measure tends to correlate slightly better with expert evaluation when k=20 and 30, compared to k=10. Therefore, for simplicity, k=20 is selected for all further LDA analysis.

Here are two example of the terms associated with topics in the fitness app task. Topic A is related to controlling food and nutrition intake and topic B is related to using social media or social network to connect with friends.

*Topic A: food, weight, nutrition, take, help, record, input, good, intake, recommend*

**Table 1. Correlation between different measures and expert evaluation.**

| Ideation Tasks | Alarm Clock App (n=200) | Fitness App (n=240) | TV Advertising (n=300) |
|---|---|---|---|
| Expert Evaluated Novelty (Mean±SD) | 4.55±1.01 | 4.64±1.38 | 3.97±1.20 |
| Correlation of LSA measure with expert evaluation | 0.114 (p=0.107) | 0.230 (p<0.001) | 0.230 (p<0.001) |
| Correlation of LDA measure (k=10) with expert evaluation | 0.184 (p=0.009) | 0.231(p<0.001) | 0.198 (p<0.001) |
| Correlation of LDA measure (k=20) with expert evaluation | 0.253 (p<0.001) | 0.190 (p=0.003) | 0.229 (p<0.001) |
| Correlation of LDA measure (k=30) with expert evaluation | 0.226 (p=0.001) | 0.219 (p<0.001) | 0.235 (p<0.001) |
| Correlation of TF-IDF measure with expert evaluation | 0.340 (p<0.001) | 0.319 (p<0.001) | 0.307 (p<0.001) |
| Correlation of crowd evaluation and expert evaluation | 0.748(p<0.001) | 0.501 (p<0.001) | 0.648 (p<0.001) |

**Table 2. Top ten novel ideas according to different measures.**

| Ideation Tasks | Alarm Clock App | Fitness App | TV Advertising |
|---|---|---|---|
| True novelty of the top ten ideas by LSA measure (Mean±SD) | 4.38±0.82 | 5.45±1.01 | 5.1±1.07 |
| Number of correctly identified top ten ideas by LSA measure | 0 | 1 | 0 |
| Point-biserial correlation comparing top 10 novel ideas by LSA measure and the remaining ideas | -0.039 (p=0.587) | 0.122 (p=0.059) | 0.176 (p=0.002) |
| True novelty of the top ten ideas in LDA measure (k=20) (Mean±SD) | 5.13±0.82 | 5.30±0.90 | 4.55±1.21 |
| Number of correctly identified top ten ideas by LDA (k=20) measure | 1 | 0 | 1 |
| Point-biserial correlation comparing top 10 novel ideas by LDA (k=20) measure and the remaining ideas | 0.132 (p=0.062) | 0.099 (p=0.125) | 0.09 (p=0.119) |
| True novelty of the top ten ideas in TF-IDF measure (Mean±SD) | 4.98±0.97 | 5.40±0.49 | 5.35±0.90 |
| Number of correctly identified top ten ideas by TF-IDF measure | 1 | 0 | 3 |
| Point-biserial correlation comparing top 10 novel ideas by TF-IDF measure and the remaining ideas | 0.098 (p=0.168) | 0.130 (p=0.045) | 0.215 (p<0.001) |
| True novelty of the top ten ideas in crowd evaluation (Mean±SD) | 5.73±0.66 | 5.9±0.44 | 5.95±0.61 |
| Number of correctly identified top ten ideas by crowd evaluation | 3 | 1 | 1 |
| Point-biserial correlation comparing top 10 novel ideas by crowd evaluation and the remaining ideas | 0.269 (p<0.001) | 0.190 (p=0.003) | 0.300 (p<0.001) |

*Topic B: app, social, challenge, friend, media, give, custom, achieve, connect, design*

As seen in Table 1, LSA measure significantly correlates with expert evaluation in two out of the three tasks. LDA measure and TF-IDF measure significantly correlate with expert evaluation for all three tasks. TF-IDF measure outperforms the other two computational methods for all three batch of ideas. Apparently, crowd evaluation correlates with expert evaluation better than all the three computational methods.

Since crowdsourcing projects commonly seek a few top ideas, we also look at the top ten novel ideas according to each measure and find out whether their true novelty (based on expert evaluation) is higher than the remaining ideas. We analyzed this by using point-biserial correlation. Specifically, we assigned a dummy variable to each idea based on whether it is a top ten

idea. Then the correlation between this dummy variable and the true novelty is calculated as the point-biserial correlation. In addition, we determined the number of correctly identified top ten ideas by each measure. This was achieved by comparing the top ten novel ideas based on expert evaluation and the top ten ideas based on each measure. When a computational measure matches expert evaluation well, the top ten ideas according to this measure would be more novel than the remaining ideas, resulting a significant point-biserial correlation. In addition, there would be much overlap between the top ten idea list based on this measure and the list based on expert evaluation. The results are in Table 2. True novelty in table 2 means the expert evaluated novelty scores. The number of correctly identified top ten novel ideas is small for all computational measures while crowd evaluation performed better in this aspect. In addition, crowd evaluation resulted in significant point-biserial correlation for all three ideation tasks. This number is 0, 1, and 2 for LDA, LSA, and TF-IDF measures, respectively.

## 5. Discussion

This study compares three computational methods for evaluating idea novelty, based on LSA, LDA and TF-IDF, respectively. An ideal computational method should correlate highly with human expert evaluation. LSA based measure correlates with expert evaluation for only two out of the three set of ideas. This is similar to the outcome in [14] where LSA distances correlated well with creativity ratings in only two out of four divergent thinking tasks.

LDA measure correlates significantly with the expert evaluation in all three tasks. This suggests that LDA measure may match human judgment more consistently than LSA measure. [7] argued that LDA is a good measure of idea similarity. Their method is also using cosine similarity between vectors of topic weights, just like our method. In their study, the correlation between LDA-based idea similarity and human judged similarity is 0.54 and 0.51 for two sub-samples. In contrast, our LDA-based novelty measure has relatively low correlation with human judgement (no more than 0.3). It is worth noting that both the number of ideas and topics are larger in [7], which might contribute to the difference. It seems worth it to explore the relationship between the validity of LDA measure (and other measures) and the number of ideas used.

Although TF-IDF is an older and less sophisticated measure among the computational methods tested, our results show it is clearly the best in matching human evaluation. It is important to recognize the advantage of TF-IDF as a simple yet useful method. At the same time, it seems surprising that LSA and LDA did not perform particularly well, given their success in many semantic modeling tasks, such as modeling semantic memory representation [13], information retrieval [12], measuring semantic similarity [11], and analyzing topics in social media [32]. There might be some limiting factors for LSA and LDA to evaluate novelty. LSA applies singular value decomposition to simplify TFIDF matrix and preserve the largest variance. Perhaps this simplification is harmful for its ability to fully identify novel idea components. LDA tries to explain how the documents are generated from the topics without fully considering whether the topics themselves are novel or not. This might limit the accuracy of novelty assessment. TFIDF is intended to detect the importance of words to a set of documents and might find the most important or innovative words which can be very indicative of idea novelty. In short, while TF-IDF shows some promise in matching expert evaluation, we agree with [14] in that the current LSA method (as well as LDA) does not correlate particularly well with human judgment in novelty evaluation.

However, even the best computational method, TF-IDF, still performs only moderately well. The correlation between TF-IDF measure and expert assessment is between 0.3 and 0.4, well below the correlations achieved by crowd evaluation. Therefore, much improvement is needed for these computational methods to resemble human evaluation.

This assertion is further supported when we look at the top novel ideas. Usually companies only care about a few top ideas in any ideation tasks. If we replace human judgment with any of the three computational methods and select top ten ideas based on the method, we would miss the majority of the true top ten novel ideas. None of the computational measures showed significant point-biserial correlation consistently across all three tasks. Consequently, if we abandon human judgment and use these computational evaluations of idea novelty, we might end up selecting ideas that are not particularly novel. In contrast, crowd evaluation is better at identifying top ideas. Therefore, even though crowd evaluation is not perfect, it is still remarkably better than the three computational methods tested. However, since LDA and TF-IDF measures are consistently significantly correlated to human judgment, they might help to eliminate many unoriginal ideas to reduce the number of candidate ideas. The accuracy of computational methods to detect unoriginal ideas is worth testing in the future.

The results in our study are aligned with [17] in that crowd evaluation is better than computational methods. Specifically, their computational method was to use LSI to calculate idea similarity so that a training set of ideas was clustered. An idea's distinctiveness (i.e. novelty)

was calculated based on the size of the cluster the idea was assigned to. It turns out distinct (i.e., novel) ideas were more likely to be implemented. However, crowd evaluation in an online community demonstrated better ability in predicting idea implementation by the company. Even though our results seem to resemble [17], there are three important differences. First, our predicted variable is idea novelty itself, not idea implementation. Therefore, our study is a direct investigation of the validity of computational novelty assessment, without introducing the confounding factor of the connection between idea novelty and idea implementation. Second, we do not need a training set of ideas to enable the computational methods. Third, we compare three computational methods based on LSA, LDA and TF-IDF, respectively.

Both our study, [14] and [17] indicate that the computational methods are not matching human evaluation particularly well. In general, these methods all adopt the bag-of-words model, treating documents as a set of words disregarding grammar, phrase, and word order. It appears that these computational methods are not able to capture all the nuances of novelty assessment by human experts, which may also include much background knowledge and associations outside the set of ideas. Another important factor is that idea novelty may have different forms. Introducing new idea elements or concepts is one form, such as introducing avatar in a mobile app. This type of novelty may be easier to detect by our computational approaches based on word frequencies. For example, avatar might be a rare word in mobile app ideas. However, there is another form of novelty: unique combination of common concepts that rarely appear together. This form of novelty might need additional methods to detect, potentially cluster analysis [24]. Yet more challenging is to evaluate the type of ideas containing common words but uncommon relationships. For example, music and vibration are commonly used features in alarm clock apps. However, it seems uncommon to turn music into vibrations. Further analysis of idea novelty and its different forms will help develop more accurate evaluation methods.

It is important to point out that our computational assessment used crowd ideas as the idea pool, instead of existing ideas on the market. Arguably the best measurement of novelty should use the ideas on the market as the reference. For example, it is possible to collect all the alarm clock apps in the market and treat their descriptions as ideas and apply our computational approaches to evaluate idea novelty. This method is well worth testing in the future work. However, we also recognize that it is not always feasible to collect all the existing ideas on the market for a specific topic. In such cases, crowd ideas, especially if the number of ideas is large, probably would cover common ways of thinking on a topic and therefore qualify to serve as a reference for novelty assessment. Therefore, the current computational methods, though not perfect, still represent a worthwhile and practical approach.

Our study suggests that much more research is needed to automate the evaluation of creative ideas. In the future, additional computational approaches should be tested on idea assessment. For example, multi-gram dictionary, if available, may be used to account for phrases [1]. Semantic network analysis has been used to evaluate creative ideas [24]. But it is complicated and demands a training set. Simplifying this approach might lead to new approaches. Additional methods using training sets to train algorithms to predict novelty scores may be fruitful as well. Having domain-specific training sets is likely to improve the accuracy of such prediction. In the meantime, the generalizability of such domain-specific training needs to be examined.

It should be noted that our methods are all about novelty assessment and do not evaluate idea usefulness, the other dimension in idea assessment. It is possible that some ideas are useful yet unoriginal. If we only focus on novelty assessment, there is a risk of filtering out these ideas. Therefore, if these computational approaches are used for practice, this limitation should be kept in mind. Conversely, what if we select ideas that are novel yet useless? It is shown that novel (but useless) stimuli or ideas can be integrated with existing useful ideas to obtain ideas that are both novel and useful [37]. Therefore, an additional use of automatic novelty assessment is to mine websites or big data for stimuli to inspire ideas that are both novel and useful [34]. Furthermore, it is highly desirable to be able to automatically select ideas that are both novel and useful. Therefore, we need to complement idea novelty assessment with idea usefulness assessment. [28] shows that idea length, specificity, readability and spelling are related to idea quality. However, these characteristics alone do not seem to be enough for identifying idea usefulness. Further improvement on automatic assessment of idea quality or usefulness would be very valuable and potentially combined with idea novelty assessment to form a complete evaluation.

With regard to practical implications, our study shows that delegating idea evaluation, or even just the first round screening, to our three computational methods, is risky. The chance of missing truly innovative ideas is significant. It is possible that our relatively small number of ideas (and number of topics in LDA) limits the performance of the computational methods. However, before further testing is done, managers of crowdsourcing projects are warned against forgoing human evaluation. Although we show some evidence that crowd evaluation is relatively reliable, it

is important to note that the accuracy of crowd evaluation is heavily dependent on domain knowledge and level of motivation. In our study, the ideas are about TV advertisement promoting public transportation to the general public, the crowd has relevant experience, knowledge, and potentially the motivation to evaluate the ideas. Such is not the case for evaluating technical ideas on designing a self-driving car. In short, our results show some promise of alternative measures of idea novelty, yet also indicate large room for improvement.

## 6. References


[1] V. Abedi, M. Yeasin, and R. Zand, "Empirical study using network of semantically related associations in bridging the knowledge gap", *Journal of Translational Medicine*, 2014, 12(1), 324.

[2] T. M. Amabile, "Social psychology of creativity: A consensual assessment technique", *Journal of Personality and Social Psychology*, 1982, 43(5), pp. 997-1013.

[3] A. Afuah, C.L. Tucci, "Crowdsourcing as a solution to distant search", *Academy of Management Review*, 2012, 37(3), pp. 355-375.

[4] B. L. Bayus, "Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community", *Management Science*, 2013, 59(1), pp. 226-244.

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

[6] D.M. Blei, "Probabilistic topic models", *Communications of the ACM*, 2012, 55(4), pp. 77–84.

[7] J. Chan and C. D. Schunn, "The importance of iteration in creative conceptual combination", *Cognition*, 2015, 145, pp. 104-115.

[8] K. J. Boudreau, E. C. Guinan, K.R. Lakhani, and C. Riedl, "Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science", *Management Science*, 2016, 62(10), pp. 2765-2783.

[9] P. Criscuolo, L. Dahlander, T. Grohsjean, and A. Salter, "Evaluating novelty: The role of panels in the selection of R&D projects", *Academy of Management Journal*, 2017, 60(2), pp. 433-460.

[10] D. L. Dean, J. M. Hender, T. L. Rodgers, and E. Santanen, "Identifying good ideas: constructs and scales for idea evaluation", *Journal of Association for Information Systems*, 2006, 7(10), pp. 646-699.

[11] K. Fu, J. Chan, J. Cagan, K. Kotovsky, C. Schunn, and K. Wood, "The meaning of "near" and "far": The impact of structuring design databases and the effect of distance of analogy on design output", *Journal of Mechanical Design*, 2013, 135. 021007

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

[13] T. L. Griffiths, M. Steyvers, and J.B. Tenenbaum, „Topics in semantic representation", *Psychological Review*, 2007, 114(2), pp. 211–244.

[14] R. W. Hass, "Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications", *Memory and Cognition*, 2017, 45(2), pp. 233-244.

[15] B. A. Hennessey and T. M. Amabile, "Creativity", *Annual Review of Psychology*, 2010, 61, pp. 569–698.

[16] J. Howe, "The rise of crowdsourcing", *Wired*, 2006, 14 (6), pp. 1–4.

[17] S. Hoornaert, M. Ballings, E. C. Malthouse, and D. Van den Poel, "Identifying new product ideas: waiting for the wisdom of the crowd or screening ideas in real time", *Journal of Product Innovation Management*, 2017, 34(5), pp. 580-597.

[18] Y. Huang, P. Vir Singh, and K. Srinivasan, "Crowdsourcing new product ideas under consumer learning", *Management Science*, 2014, 60(9), pp. 2138-2159.

[19] H. J. Jung, and J.J. Lee, "The quest for originality: A new typology of knowledge search and breakthrough inventions", *Academy of Management Journal*, 2016, 59(5), pp. 1725-1753.

[20] S. Kaplan, and K. Vakili, "The double-edged sword of recombination in breakthrough innovation", *Strategic Management Journal*, 2015, 36(10), pp. 1435-1457.

[21] M. Klein, and A. C. B. Garcia, "High-speed idea filtering with the bag of lemons", *Decision Support Systems*, 2015, 78, pp. 39-50.

[22] T. K. Landauer, and S.T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge", *Psychological Review*, 1997, 104, pp. 211–240.

[23] H. Lee, and S. Seo, "What Determines an Agreeable and Adoptable Idea? A Study of User Ideas on MyStarbucksIdea. com", 2013 46th Hawaii International Conference on In System Sciences (HICSS), 2013, pp. 3207-3217.

[24] M. D. R. Mart ínez-Torres, F. Rodriguez-Pi ñero, and S. L. Toral, "Customer preferences versus managerial decision-making in open innovation communities: the case of Starbucks", *Technology Analysis & Strategic Management*, 2015, 27(10), pp. 1226-1238.

[25] N. Peladeau, and E. Davoodi, "Comparison of Latent Dirichlet Modeling and Factor Analysis for Topic Extraction: A Lesson of History", *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[26] H. Piezunka, and L. Dahlander, "Distant search, narrow attention: How crowding alters organizations' filtering of suggestions in crowdsourcing", *Academy of Management Journal*, 2015, 58(3), pp. 856-880.

[27] C. Riedl, I. Blohm, J. M. Leimeister, and H. Krcmar, "The effect of rating scales on decision quality and user attitudes in online innovation communities", *International Journal of Electronic Commerce*, 2013, 17(3), pp. 7-36.

[28] M. Rhyn and I. Blohm, "A Machine Learning Approach for Classifying Textual Data in Crowdsourcing", in Leimeister, J.M.; Brenner, W. (Hrsg.): *Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik* (WI 2017), St. Gallen, S. 1171-1185.

[29] G. Salton, and C. Buckley, "Term Weighting Approaches in Automatic Text Retrieval," *Information Processing & Management*, 1988, 24(5), pp. 513–523.

[30] B. Schemmann, A. M. Herrmann, M. M. Chappin, and G. J. Heimeriks, "Crowdsourcing ideas: Involving ordinary users in the ideation phase of new product development", *Research Policy*, 2016, 45(6), pp.1145-1154.

[31] Z. Shi, G. M. Lee, and A. B. Whinston, "Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence", *MIS Quarterly*, 2016, 40(4), pp. 1035-1053.

[32] A. H. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, L. H. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach", *PLoS ONE*, 2013, 8(9), e73791.

[33] O. Toubia, and O. Netzer, "Idea Generation, Creativity, and Prototypicality", *Marketing Science*, 2017, 36(1), pp. 1-20.

[34] K. Wang, "Collective innovation: A literature review", In Technology Management in the IT-Driven Services, 2013 *Proceedings of PICMET*' (pp. 608-615). IEEE.

[35] K. Wang, and J.V. Nickerson, "A literature review on individual creativity support systems", *Computers in Human Behavior*, 2017, *74*, pp. 139-151.

[36] K. Wang, J. V. Nickerson, and Y. Sakamoto, "Crowdsourced Idea Generation: The Effect of Exposure to an Original Idea", *Creativity and Innovation Management*, 2018, 27 (2), pp. 196-208.

[37] K. Wang, H. Wang, and Y. Tao, "Combining Ideas in Crowdsourced Idea Generation", *Foundations of Management*, 2017, 9(1), pp. 203-212.

[38] T. P. Walter, and A. Back, "A text mining approach to evaluate submissions to crowdsourcing contests", *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2013, pp. 3109–18.

[39] A. Westerski, T. Dalamagas, and C. A. Iglesias "Classifying and comparing community innovation in idea management systems", *Decision Support Systems*, 2013, 54 (3), pp. 1316–26.