Exploring Machine-based Idea Landscapes – The Impact of Granularity

Julian Wahl
University of Innsbruck
julian.wahl@uibk.ac.at

Thomas Ströhle
University of Innsbruck
thomas.stroehle@uibk.ac.at

Johann Füller
University of Innsbruck
johann.fueller@uibk.ac.at

Katja Hutter
University of Innsbruck
katja.hutter@uibk.ac.at

Abstract

Effective exploration of a landscape full of crowdsourced ideas depends on the right search strategy, as well as the level of granularity in the representation. To categorize similar ideas on different granularity levels modern natural language processing methods and clustering algorithms can be usefully applied. However, the value of machine-based categorizations is dependent on their comprehensibility and coherence with human similarity perceptions. We find that machine-based and human similarity allocations are more likely to converge when comparing ideas across more distant solution clusters than within closely related ones. Our exploratory study contributes to research on the navigability of idea landscapes, by pointing out the impact of granularity on the exploration of crowdsourced knowledge. For practitioners, we provide insights on how to organize the search for the best possible solutions and control the cognitive demand of searchers.

1. Introduction

In crowdsourcing contests, innovation-seeking organizations reach out to a broad range of people with distant and diverse perspectives to gather possible solutions to their problems [1, 20, 50]. To explore not only one idea but a whole idea landscape and thereby obtain a better understanding of potential solutions, it is important to discover the broadest possible opportunity space [47, 55]. Through the sourcing of diverse perspectives in crowds, rich landscapes full of solution-related knowledge can be searched [18, 20]. The insights gained there may help to find the best overall solutions. More precisely, the emerging idea landscapes give an overview about the number of possible solutions, how distant or close solutions paths are to each other, which ideas are positioned in the near neighborhood, the size of solution clusters with similar ideas, as well as insights into the problem structures. Numerous possible ways of solving a problem can be recognized which is quite challenging, especially when facing a high diversity of ideas and different problem structures.

Innovation research using a landscape metaphor suggests that knowledge searches are most efficient if they are performed in successive patterns [24, 26, 44]. While a broad overview over various areas of a landscape may help decision-makers to gain insights into the general structure of themes, at a certain point they are well-advised to focus their search on areas identified as particularly promising [47]. Thus, to effectively explore spaces full of crowdsourced knowledge, it is not only the right search strategy [33] but also an accurate representation of the landscape's structure at the right resolution that matters [11, 37].

Categorizations on different granularity levels may be beneficial to explore the extensive sets of ideas in an efficient and effective way. Thus, when designing and representing idea landscapes it is essential to apply the right granularity and categorize the diverse but potentially overlapping sets of ideas into solution clusters that ensure human comprehension [2, 37]. A too narrow landscape may cause huge search efforts and lead to a "do not see the wood for the trees" effect. In contrast, a landscape being too rough may be too superficial to discover important details. Related to our research context, this means that it is not obvious how an appropriate level of granularity between single ideas or groups of ideas should look like. Additionally, the determination of nuanced differences in similarity perception and coherent categorizations has turned out as quite challenging and resource-intensive for humans [13, 16, 29, 54].

The idea of representing large pools of possible solutions as landscapes is an exciting topic in theory but until recently the creation of such representations

URI: https://hdl.handle.net/10125/79372
978-0-9981331-5-7

has been considered as a rather tedious work that involves a substantial degree of human effort and time [29]. Recent studies in the field of innovation search and landscape exploration suggest that the natural language processing (NLP) methods of word and document embeddings are promising to reduce the effort of landscape creation and exploration based on a semantic similarity allocation of ideas [17, 27, 32]. Document embeddings allow to compare text documents based on their word similarities and represent them as numeric vectors at specific locations in an embedded space depending on their semantic meaning [7, 30, 31, 45].

In times where ideas can be easily mapped with the support of text mining and powerful NLP methods, still little is known how to best represent ideas and their similarities in a way that best supports innovators in exploring the best ones. When ideas are categorized in incoherent ways, innovators may struggle with incomprehension and experience increased cognitive load [11, 37, 49]. Consequently, the merit of machine-based categorizations depends on their comprehensibility and coherence with those generated by humans across various granularity levels.

Thus, in our exploratory study, we are not only interested in how the extensive and diverse knowledge shared in a crowdsourcing contest can be structured into meaningful representations with the help of modern NLP methods and clustering algorithms but also want to find out how different granularity levels affect the information processing of humans and the navigation in machine-based idea landscapes in further consequence.

We find that the applied granularity level, indeed, plays an important role. Human similarity perceptions are more likely to comply with a machine-based allocation when distinguishing between ideas across coarser granularity levels than on finer levels. With our study, we contribute to research on the navigability of ideas landscapes [10, 32, 52, 53] by showing that human similarity perceptions are better matched on coarser cluster granularity levels. This has important implications on how to navigate through idea landscapes and control the cognitive demand of searchers when exploring ideas. Furthermore, we also illustrate how modern NLP methods for defining semantic similarities can structure and analyze myriads of solution-related knowledge shared in crowdsourcing contests more efficiently and effectively [28, 42].

2. Literature background

2.1. Idea landscapes

Over the last several years, crowdsourcing contests have gained momentum to solicit novel solutions to innovation problems from external and internal sources through online platforms [12, 20, 50]. Compared to situations where organizations or teams search by themselves, through crowdsourcing richer landscapes of potential solutions can be accessed [18]. The participants have information about different parts of the need and/or solution landscape [18] and share their diverse perspectives on the problem [20]. The submitted ideas are discrete descriptions of potential solutions that comprise a specific configuration of features [10] and can be seen as opportunities to create value through investment [14, 29]. They take positions in an opportunity space that incorporates a collective representation of the solution-related knowledge of the crowd.

Landscapes offer a valuable metaphor to think about the space inventors need to search when pursuing new economic opportunities [2, 10, 18]. They can be defined as data representation spaces that abstract information gathered about entities in a search space and reveal interrelations among them based on their features [25, 46]. As idea features can be shaped and (re-)combined [10, 15], theoretically innumerable, if not indefinite, opportunities for entrepreneurial and innovative activity can be created [9].

Problem-solving research suggests that to come up with the best possible solution, solvers should not limit themselves to the analysis of single search paths but rather consider a broad collection of possible solutions [47]. The generation of multiple ideas and designs is not only beneficial as they may be better than the previous ones but also because they facilitate a new way of thinking [55]. Every idea shared in a crowdsourcing contest may involve valuable information for developing further useful solutions to an innovation problem in the next step [34]. However, in practice, when organizations review ideas they often focus on filtering out a specific number of winning ideas, as reading through all submissions and manually structuring the crowdsourced knowledge requires extensive resources in terms of time, money, and cognitive effort [4, 35]. As a result, a myriad of perspectives is not considered.

In situations where one is confronted with a high load of information, heuristics provide mental shortcuts that help to reduce the cognitive load in problem-solving [23, 49]. As problem solvers, innovators can pursue different search heuristics in exploring extensive landscapes full of knowledge and

opportunities [33]. Aggregated and structured information on solution-related knowledge provides valuable insights that can change the overall problem representation [19, 40, 46] and help to decide on which areas to put more focus on. Thus, when applying heuristics for idea exploration, nuanced insights into crowdsourced knowledge structures are important.

While in innovation research landscape metaphors have often been used to analyze the configuration of patents in a technology landscape that inventors search [2, 10, 32], Kornish and Ulrich [29] focused on the structural elements of sets of crowdsourced ideas. In their analysis, they focused on the size of the spaces, the redundancy of ideas, and the allocation of ideas determining the landscape structure. Therefore, ideas were manually clustered into appropriate subclasses of shared needs and categories based on their semantic similarity. Such insights into the landscape structure are useful for idea exploration, as landscape searches are most efficient if they are performed in successive patterns combining knowledge breadth and depth [24, 26, 44]. For example, at the beginning of an innovative search, individuals, as well as organizations are advised to discover areas far away in the landscape, but at a certain point – e.g. when options for technological improvement have been identified as particularly interesting - they should focus their search on local areas. A broad overview of various areas of a landscape may help decision-makers to gain insight into the general structure of themes or neighborhoods. In this context, Kornish and Ulrich [29] empirically showed that the quality of the generated ideas is higher when located in densely populated areas of the idea landscape. Their findings suggest that agglomeration patterns may not only mirror higher innovation activity around certain areas in the idea landscape [2] but can be indicative of valuable knowledge pools. A fine-tuning in selected areas may further improve the understanding of cluster-specific knowledge as well as the ease with which it can be integrated into the existing knowledge stock to find the best solutions [34].

2.2. Landscape similarity

To create meaningful idea landscapes and identify knowledge structures that enable faster and more efficient searches interrelations between ideas need to be revealed [25, 46]. While in previous studies on innovation search, the identification of interrelations among ideas in a landscape was often based on manual patent categorizations [2, 10] or resource-intensive human similarity assessments [29], today, the advancements in text mining and NLP allow for automated mapping of idea texts according to their semantic similarity. For example, to complement the rather rigid patent categorizations, Lee et al. [32] relied on word embeddings to construct a product landscape as a vector space locating similar technologies close to each other and to identify product areas with configurations of interest. The method of word embeddings allows to represent words as numeric vectors and map them according to their semantic similarity [36, 41]. Previously, text mining and NLP have also been applied to structure crowdsourced ideas into spatial representations. For example, Toubia and Netzer [51] used semantic networks to analyze the structure of a large pool of ideas. Other studies relied on topic modeling algorithms to reveal latent themes in crowdsourced idea descriptions to support the search through the solution-related knowledge [6, 22, 28] and even compared it to human similarity perceptions [53]. All these approaches share the idea of representing texts based on their semantic similarity. Importantly, the value of machine-based similarity allocation of ideas is dependent on its ability to create categorizations that are comprehensible and comply with those generated by humans [53]. Thus, some more insights into the processes of human similarity perception may be of value.

According to the theoretical concept of Tversky [54], an object's similarity – defined as a proximity relation between two objects represented by their features and properties – is modified by the individual classification in a human's mind. Cognitive research further suggests that the underlying process in similarity comparisons is one of a structural alignment or matching of two mental representations aiming for maximum coherence [13]. However, as mental representations of individuals differ, in many cases, there is more than one structural consistent match. For example, a black smartphone may be either perceived more similar to an object of the group of black objects (emphasizing a design feature) or an object of the group of communication devices (emphasizing a functional feature) depending on individual structural alignment in the similarity assessment. Not surprisingly, research on innovation and idea similarity observed that human similarity perceptions are highly sensitive when similarity assessments are broken down into certain categories and, thus, regarded overall similarity assessments as most feasible [16, 29, 54].

Both, manual similarity allocations and those created by text mining and NLP algorithms are based on structural regularities and feature similarities. While NLP algorithms determine the similarity of ideas by statistically analyzing text features, humans

may apply various feature criteria for categorizing ideas. However, to explore possible solutions in a fast and efficient way, machine-based idea landscapes must be structured into meaningful and robust similarity allocations understandable for human cognition. At least, literature seems to agree that there is something like similarity hierarchies that summarize ideas on more general levels depending on the perceived structural similarity [43], e.g. unrelated, related or identical objects. Quite often the similarities between objects are represented as hierarchical knowledge structures based on categorizations and distance-based idea clustering [13, 16, 29, 54].

2.3. Landscape granularity

In the field of geography - the science of space - the term granularity either expresses the fineness of semantic objects classed in a hierarchy or the spatial resolution of a landscape in a map by defining which entities become indistinguishable [48]. Research on geographical spaces suggests that human perception and interpretation of the properties of the space depends on the applied scale in the representation [37]. Deviations and ambiguities in the semantic granularity levels may affect the cognitive burden and perceived coherence within groups of ideas and the usefulness of the knowledge representation [11, 37].

To navigate with maps there is no need to display every detail of a landscape. In geography, the abstraction of mapped data is described as a process called amalgamation process in which previously separated features of a map are merged into indistinguishable entities from a detailed representation into a coarser one [48]. For example, when someone wants to know more about all countries eligible for a trip to Europe, the mapping of all streets connecting any rural town is unnecessary and counterproductive. However, after deciding on a set of countries, one may benefit from a different resolution scale to navigate to points of interest within certain European countries or regions. Innovation researchers can ask similar questions about optimal granularity when exploring product ideas. For example, when searching for new communication devices, the color of each device provides limited information value. However, at first, insights into the supported cellular network standard may be more relevant and only later design features like color or shape come into consideration. Thus, analogous to the amalgamation of map features, ideas can be structured into indistinguishable groups, categories, or themes in a hierarchical representation of the crowdsourced knowledge. Thereby, it has to be decided when ideas are summarized into common categories. While a red telephone, a black smartphone, and a green smartwatch can be usefully amalgamated into communication devices, on a finer granularity level the categorizations might be less clear. For example, the two latter ones could be summarized into devices facilitating 4G cellular network standard but on the same granularity level, the former two could also be categorized into devices that need to be held in hands.

Following on section 2.2., human similarity perceptions may not only differ depending on the selected feature criteria but also on the level of analysis applied for the categorization. Consequently, they can be considered as an essential factor that influences the appropriate level of granularity of an idea landscape. In other words, an accurate representation of granularity levels in machine-based idea clusters that complies with human categorizations is important to facilitate search heuristics that reduce cognitive load and enable effective idea exploration [33, 43].

To sum up, the extensive and diverse set of ideas shared in crowdsourcing contests are an ideal example of a rich landscape full of opportunities and knowledge. As the value of the machine-based representation is closely related to mental similarity categorizations and semantic knowledge hierarchies, it is important to know more about appropriate granularity levels to effectively navigate through the idea content [53]. When ideas are amalgamated in incoherent ways, humans may struggle with incomprehension, confusion, and increased cognitive load [11, 37, 49] undermining the value of machine-based similarity allocations for idea exploration. In times where idea texts can be easily mapped and structured with the help of NLP methods, we should deal with this challenge all the more. At the moment, we still know too little about how to best represent and categorize ideas.

3. Methodology

In our exploratory study, we aim to not only structure the crowdsourced knowledge using text mining and NLP but also want to know more on how to choose the granularity levels of an idea landscape to make the exploration, analysis, and finding of the best possible solutions as smooth as possible. Thus, our study is organized into four steps. In the first step, we retrieved an extensive set of ideas submitted to a crowdsourcing contest through web scraping and text mining to accumulate a sufficiently large knowledge base of possible solutions. In the second step, we structured the myriads of ideas into a machine-based idea landscape. Therefore, we applied a pre-trained word embedding and transformed it into tf-idf

weighted document embedding representing the crowdsourced ideas. In a third step, we generated a similarity hierarchy to reveal highly interrelated ideas using a hierarchical clustering algorithm. This allowed us to structure the idea vectors into idea landscapes of three different granularity levels. Finally, to find out to what extent machines categorize in the same way as humans and how information processing varies across different granularity levels, in a similarity experiment we compared the vectorized idea allocation to human similarity assessments.

3.1. Data

On the OpenIDEO platform, different initiators – ranging from governmental organizations to private firms and NGOs – can host crowdsourcing contests that propose a specific problem to be solved by more than 17,000 users from over 170 different countries. The contest "Circular Design Challenge" tackled the question of how to get products to people without generating plastic waste. Over six months, 483 participants submitted 619 ideas to the innovation contest. Participation was open to everybody and winning ideas were rewarded with monetary incentives. We scraped all the contest data from the platform using a web crawler, including all public data about participating users, data about each submitted idea. For our analysis, we only used the text of the idea descriptions. The final corpus of all 619 ideas contributed to the innovation contest accumulates to a total of 219,244 words. In an automated text cleaning step of the ideas, differences in the text structure within the ideas were eliminated. The text corpora were tokenized into unigrams. Stop words, punctuations and digits were removed [3]. All computations were performed using the statistical programming language R.

3.2. Idea vectorization

In many NLP tasks applications of word embeddings, such as word2vec [36] or GloVe [41] are widely spread. They allow for a representation of words with similar meanings in the form of numeric vectors. Word embeddings can be advanced to document embeddings which represent an unsupervised method for learning distributed representations for longer pieces of texts [7, 30, 31, 45]. Thereby, semantically similar texts are located close to each other in an embedded space. For example, the sentence "Gave an innovation talk in Maui" must have a similar semantic vector representation as "Had a new product development lecture in Honolulu". Research has shown that the representation of documents based on word vector models outperforms other popular models for semantic document representation such as tf-idf, LDA, or LSI in document similarity classification tasks [7, 30].

As ideas in crowdsourcing contests are described as texts, document embeddings seem to be a useful approach to determine the similarity of ideas based on word features and interrelations among them. Every web-scraped idea description can be transformed into an idea vector taking a specific position in the idea landscape. The idea vectors consist of numeric features that configure the semantic meaning of an idea.

To create idea vectors we applied the pre-trained word vector model "Gensim Continuous Skipgram" [38] based on Word2Vec [36] as a complementary data source.[1] The language model was trained on an English Wikipedia Dump and a Gigaword dataset corresponding to a lemmatized vocabulary of about 260 thousand words. A pre-trained embedding has the advantage that no new neural network has to be trained, which is computationally faster [38]. Furthermore, the application of pre-trained embeddings is well suited to reliably represent documents of smaller size as the calculation of idea vectors is independent of the size of the word distribution across the documents in the dataset. The vocabulary size of the pre-trained embedding was large enough for a meaningful intersection with the vocabulary of the crowdsourced ideas and only proper names or misspelled words were dropped from the dataset.

For the representation of ideas, we applied an approach for generating document embeddings that turned out to outperform other popular ones in representing shorter lengths of user-generated content [45], e.g. reviews or single ideas. Thereby, we implemented a tf-idf weighted average to summarize several word vectors into single idea vectors. The tf-idf index allows to weight each word of an idea according to its relative importance across documents [21]. However, in our case, it is not used to sort out words in single ideas but to usefully weight the words as idea components according to their importance. Finally, we applied a principal component analysis to reduce the dimension of idea vectors collected as a matrix to its first 100 principal components [45]. To

[1] We also tested document embeddings based on pre-trained GloVe and FastText models which indicated similar results.

measure the distances between the idea vectors we use the cosine similarity, which is given by the Euclidean dot product.

3.3. Idea clustering

In the next step, we measured similarities distances between ideas. These distances were then used to find out how machines amalgamate ideas into different granularity levels and knowledge hierarchies. Hierarchical clustering is a well-established contrast model to represents object similarities based on common and distinctive features [54]. Through a hierarchical cluster algorithm [39] the crowdsourced ideas were categorized into homogeneous groups using the cosine similarity between idea vectors as a metric and a complete linkage method. In our analysis, we were not interested in finding the optimal number of clusters, but rather in the different clusters at different cosine distances. Compared to other methods the advantage is that clusters can be formed at any granularity of cosine distance. The arrangement of the hierarchical clusters is illustrated in a dendrogram, which is a tree with branches indicating the cluster assignment of similar idea vectors (Figure 1). It represents a hierarchical knowledge structure of all potential solutions gathered in the crowdsourcing contest.

To find out whether the idea vectors can be organized into meaningful similarity hierarchies we concentrated on two different abstraction levels of the idea landscape. At each level, we determined clusters by cutting the dendrogram tree.[2] When the cut was made at a low height we defined the subordinate branches as a local cluster describing shared needs, when it was made higher in the dendrogram as a global cluster describing shared categories. Thus, out of the initial landscape of 619 ideas, we created another two landscape representations at different granularity levels – a finer one with 265 solution clusters or shared needs containing about 2-5 ideas and a coarser one with 20 solution clusters or shared categories of about 8-16 ideas.

To simplify the representation and facilitate human interpretation of the idea landscape we transformed the distances of the idea vectors with 100 dimensions into a two-dimensional Cartesian space through Kruskal's non-metric multidimensional scaling based on pairwise cosine similarity distances [5]. In Figure 2 a cutout of landscape in a two-dimensional space is illustrated. While the green and red colors in the graph describe the two global clusters, shape symbols are assigned to distinguish between local clusters. For example, at the bottom part of the cutout, there is a local cluster (579, 180) dealing with the mechanical design of straws and wrappers and at the center-left, there are two closely related local clusters (190, 62, and 544) proposing the usage of bottle caps as a plaything for children. In the other global cluster at the top of the cutout, there is a local cluster (344, 108, 383) discussing the conversion of plastic into oil and at the center, another local cluster (182, 189, 87, 364, 275) ideas around bioplastics can be spotted. Ideas dealing with bottle caps and the mechanical design of straws can be summarized into the global cluster category of bottle toppers. Ideas about the chemical conversion of plastic to oil and bioplastics can be meaningfully summarized into the global cluster category of chemical transformations.

In Table 1 we provide more insights into the content of the idea clusters at the different granularity levels by showing the most relevant words according to the tf-idf index weight which served as the main parameter for creating the idea vectors. All of the illustrated ideas on the idea level (344, 108, 383) discuss the conversion of plastic into oil and, thus, build a local cluster in the landscape describing shared needs. At a coarser granularity level, this local cluster itself is a part of a global cluster focusing on chemical transformations representing one of the 20 solution clusters in the landscape describing shared categories.

Figure 1. Dendrogram cutout of global and local solution clusters

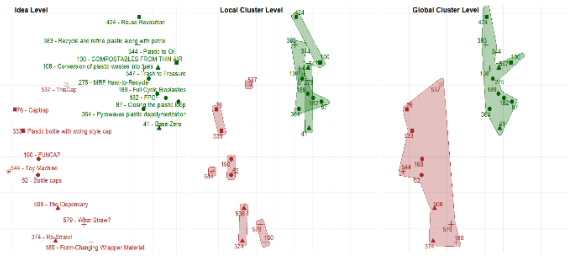

Figure 2. Cutout of the idea landscape at three landscape granularities[3]

[2] To identify shared needs and shared categories we orientated to the procedure of Kornish and Ulrich [29]. We slightly adapted the two cut-off values due to specific characteristics our idea set.

[3] The projection from a 100-dimensional to a two-dimensional space can lead to an overlapping of clusters in the two-dimensional space.

Table 1. Top-10 words according to tf-idf weight at three landscape granularities

"Conversion of plastic wastes into fuels"	"Recycle and refine plastic along with petrol"	"Plastic to Oil"	"Conversion of plastic into oil"	"Chemical transformations"
108 - Idea Level	383 - Idea Level	344 - Idea Level	Local Level	Global Level
elp	petroleum	fuel	fuel	pha
fuel	refinery	anaerobic	elp	biocellection
synthetic	oil	asphalt	refinery	pod
landfill	filter	bunker	petroleum	feedstock
technology	gasoline	coastline	oil	fpc
cauterize	process	lubricant	filter	greenwaste
colossal	watch	myriad	gasoline	bioplastic
gyre	gas	ton	synthetic	jose
illegally	invention	paraffin	watch	conversion
obstruct	country	roughly	invention	san

Importantly, the hierarchical clustering allowed us to abstract the initial allocation on the idea level into various granularity levels. While our results support previous findings suggesting that ideas of large and diverse knowledge sources tend to cluster towards semantically similar themes [2, 29], the analysis also offers nuanced insights into the shared idea content at different generalization levels by looking at prominent words on individual granularity levels.

To test the numerous cluster solutions at different granularities for their interpretability and face validity, we roamed through all global and local clusters in the landscape. After this first qualitative analysis of the similarity allocation, we could confirm the applicability of document embeddings to create meaningful similarity hierarchies and categorizations that may facilitate effective navigation through idea landscapes. However, in our analytical search - focusing on the meaningfulness of the allocations rather than the identification of interesting ideas - we also had the impression that the differentiation between idea categories was easier across global cluster levels than within them across local ones. In the subsequent similarity assessment task, we empirically checked this impression to extend our knowledge about the impact of granularity on human similarity perception.

3.4. Similarity assessment

After revealing the hierarchical knowledge structure with the help of a machine-based idea similarity allocation, we now focus on how the human similarity perception relates to machine-based ones and to what extent the categorizations diverge on different granularity levels. Thus, we set up a web-based similarity assessment consisting of two task types. In both task types, participants had to identify a pair of the two most similar ideas out of three given ideas [53].

In the global comparison task, we tested whether the human assessments comply with the referenced cluster solutions in the landscape across coarser granularity levels. Therefore, one idea pair from a local cluster and a third idea from a different global cluster were randomly drawn. Consequently, each global comparison task consisted of two ideas close to each other and one idea further away in the idea landscape. In a local comparison task, we tested whether the human assessments comply with the referenced cluster solutions in the landscape across coarser granularity levels. Therefore, one idea pair from a local cluster and a third idea from the same global cluster were randomly drawn. Consequently, each local comparison task consisted of three ideas with a relatively high level of semantic similarity located in neighborhoods. We hypothesize that this task category may be more challenging than the global one. Further, it is notable that even within the set of local comparison tasks participants may encounter various degrees of difficulty due to differences in the density of similar ideas in the clusters.

In total 34 participants with sufficient English proficiency to understand the idea content took part in the human similarity assessment. Each participant had to complete five comparison tasks. For each task, they were provided with a subset of three idea descriptions. Then the participants were asked to identify a pair of most similar ideas out of a subset. We intentionally instructed them to use their own notion of overall similarity for completing the comparison task [16, 29].[4] We collected 64 scores of the local and 56 scores of the global comparison task, resulting in a total of 120 human assessments scores, which were compared to the machine-based similarity allocations.

The results of the similarity assessment task show that in 76.8% of the global tasks the human similarity assessment followed the similarity allocation in the idea landscape. Concerning the local tasks, the human similarity perception converged to the machine-based allocation in 59.4% of the cases. For both task types, the success rate for recognizing the semantically more similar pairs is clearly distinguished from a random selection or guessed response to the experiment which would be 33.3%.

Based on these results we find that the level of semantic granularity in the landscape, indeed, plays an

[4] One of the tasks was an attention check where we included two similar ideas and an unrelated idea from another contest. Four people were sorted out due to wrong answers.

important role when representing crowdsourced ideas in a machine-based idea landscape. Overall, it is striking that machine-based semantic similarities can also be recognized similarly by humans. However, the results of the similarity assessment task also indicated that the machine-based idea similarities and human similarity perceptions are more likely to converge when comparing ideas across coarser granularity levels in global comparison tasks than on finer levels in local comparison tasks. Looking at it from a human perspective the machine-based allocation seems more intuitive and accurate when exploring coarser knowledge structures across global clusters. However, the approach is still fairly useful to distinguish between ideas across fine granularity levels within global clusters. These findings are important to confirm the applicability and value of machine-based idea landscapes.

4. Discussion

With this study, we contribute to research on the navigability of ideas landscapes [10, 32, 52, 53]. A smooth and effective exploration of ideas requires sufficient matching between machine-based and human similarity categorizations on different hierarchical levels. Our findings suggest that human similarity perceptions are more likely to converge with the machine-based allocation on coarser granularity levels. The lower agreement between the similarity allocations of humans and the machine on finer granularity levels may be attributed to various possible reasons which are worth a closer look.

From a human's point of view, it seems plausible that the individually different mental representations have a stronger effect on finer granularity levels yielding lower agreement rates. Idea features may be interpreted differently and lead to many different but structural consistent idea similarity assessments in human information processing systems [13]. Humans are likely to perceive a higher cognitive load [49] in processing three highly similar ideas on local levels which negatively affects the accuracy of their similarity assessments. However, it is also possible that the machine-based similarity allocation loses accuracy on finer granularity levels. The differences in similarity processing of humans and machines may deliver a more balanced explanation. While similarity perceptions of humans often rely on the interpretations of idea features based on their learnings, experience, or social backgrounds [13, 19], machines learn their similarity representation through a purely statistical analysis of regularities within large sets of letters and words. To find out which of the proposed explanations is valid, more research on human and machine-based similarity allocation and perception is needed.

In the course of the study, we also illustrate how modern NLP methods for defining semantic similarities can structure and analyze myriads of solution-related knowledge shared in crowdsourcing contests in an efficient and effective way. In particular, we have shown how an extensive set of crowdsourced ideas can be structured into meaningful knowledge hierarchies. The approach frees up valuable resources such as time, costs, and manpower. A comparison to the study of Kornish and Ulrich [29] who analyzed the structure of an idea landscape in a related setting illustrates the potential benefits. In their study, they manually created a semantic similarity clustering of 400 ideas by engaging 230 human raters. They needed between 30-50 minutes to complete a grouping task and $ 10 was paid as compensation to each rater, adding up to around 9200 minutes of time effort and costs of $ 4000. We generated an adequate clustering of 619 ideas in just a few seconds without any noteworthy costs. This exemplifies how NLP methods like document embeddings can substitute tedious tasks like reading through every idea, meaningfully organize the diverging perspectives shared by the crowd and dramatically reduce human effort [28, 32, 42]. The efficient and effective representation of crowdsourced ideas facilitates innovation research to learn more about the size and structure of landscapes full of possible solutions, how distant or close solutions are to each other and the properties of individual landscapes at different granularity levels [29]. Previous research suggests that dense agglomerations in a landscape create an unbiased structure that is indicative of higher innovation activity and ideas addressing a relevant problem or need [2, 29]. Nevertheless, smaller clusters might also be interesting to research as they entail relatively unique solutions. While our research focused on ideas generated in a crowdsourcing contest, the findings on the impact of granularity in landscape generation are also relevant for other settings where knowledge is explored to identify the best possible solutions, such as allocations of patents or design concepts [32, 56].

For innovation managers, the segmentation of the idea landscape into different knowledge hierarchies and solution clusters offers useful functions that support the search for the best possible solution. By selecting appropriate granularity levels, innovation managers can get a fast and comprehensive picture of the solution-related knowledge without reading through every idea. Depending on the selected granularity levels broader or narrower parts of the landscape can be explored. Our findings suggest that for humans, machine-based idea categorizations are

more intuitive and agreeable on broader than finer levels. Thus, innovation managers are advised to apply search heuristics [33, 47] that start with the exploration of broader solution clusters to avoid cognitive load due to different similarity perceptions and dive deep into finer solution clusters at a later stage to ensure the depth of the search [24, 26, 44, 47]. Hereby, NLP methods enable an automated mapping of idea content on different abstraction levels to obtain a fast glimpse of important themes. The idea clusters can be further combined with aggregated information, e.g. the number of ideas, likes, or sentiments in comments which may help to decide on which areas to put more focus on. However, to create new combinations and reconfigurations [10, 44] based on the exploration of the crowdsourced knowledge in an idea landscape, human cognition involving in-depth knowledge and experience about the organizational context, market situations, or competitive scenarios may remain indispensable to assess the values of single ideas or groups of ideas. The semantic allocation of possible solutions should support innovation managers to abstract idea content and facilitate an efficient and effective discovery of new ideas. Other scholars are encouraged to build on our work to find out more on how the diverse knowledge sources in the different parts of an idea landscape should be combined to build a holistic understanding about possible solutions and identify the best overall solutions.

Furthermore, we are aware that the applied document embedding method is only one possibility to reduce ideas to their essentials and to measure their semantic distances with idea vectors. Different embeddings result in subtle differences in the semantic allocation and clusters affiliation, but also different methods of distance calculation may lead to changes in results. In the future, it could be interesting to compare different document embeddings, including contextualized language models such as BERT [8], and evaluate which methods work best to differentiate between crowdsourced ideas in the given conditions. While our work measured similarity perception on only two granularity levels, increasing the number of granularity levels could also help to find out more about the optimal way to represent idea landscapes.

6. References

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

[2] Aharonson, B.S., and M.A. Schilling, "Mapping the technological landscape: Measuring technology distance, technological footprints, and technology evolution", Research Policy 45(1), 2016, pp. 81–96.

[3] Arnold, T., "A Tidy data model for natural language processing using cleanNLP", R Journal 9(2), 2017, pp. 248–267.

[4] Blohm, I., C. Riedl, J. Füller, and J.M. Leimeister, "Rate or trade? Identifying winning ideas in open idea sourcing", Information Systems Research 27(1), 2016, pp. 27–48.

[5] Cox, M.A.A., and T.F. Cox, "Multidimensional Scaling", In Handbook of Data Visualization. Springer, Berlin, Heidelberg, 2008, 315–347.

[6] Cui, T., and L. Liu, "Identifying Successful Ideas in Crowdsourcing Contest: Effects of Idea Content and Competition Intensity", Proceedings of the 41st ICIS, (2020).

[7] Dai, A.M., C. Olah, and Q. V. Le, "Document Embedding with Paragraph Vectors", 2015, pp. 1–8.

[8] Devlin, J., M.W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[9] Felin, T., S. Kauffman, R. Koppl, and G. Longo, "Economic opportunity and evolution: Beyond landscapes and bounded rationality", Strategic Entrepreneurship Journal 8, 2014, pp. 269–282.

[10] Fleming, L., and O. Sorenson, "Science as a map in technological search", Strategic Management Journal 25(89), 2004, pp. 909–928.

[11] Fonseca, F., M. Egenhofer, C. Davis, and G. Câmara, "Semantic granularity in ontology-driven geographic information systems", Annals of Mathematics and Artificial Intelligence 36(1–2), 2002, pp. 121–151.

[12] Füller, J., K. Hutter, J. Hautz, and K. Matzler, "User roles and contributions in innovation-contest communities", Journal of Management Information Systems 31(1), 2014, pp. 273–308.

[13] Gentner, D., and A.B. Markman, "Structural Alignment in Comparison : No Difference without Similarity", 5(3), 1994, pp. 152–158.

[14] Girotra, K., C. Terwiesch, and K.T. Ulrich, "Idea generation and the quality of the best idea", Management Science 56(4), 2010, pp. 591–605.

[15] Goldenberg, J., D. Mazursky, and S. Solomon, "The fundamental templates of quality ads", Marketing Science 18(3), 1999, pp. 333–351.

[16] Griffin, A., and J.R. Hauser, "The Voice of the Customer", Maketing Science 12(1), 1993, pp. 1–27.

[17] von Hippel, E., and S. Kaulartz, "Next-generation consumer innovation search: Identifying early-stage need-solution pairs on the web", Research Policy 50(8), 2021.

[18] von Hippel, E., and G. von Krogh, "Identifying Viable 'Need-Solution Pairs': Problem Solving Without Problem Formulation", Organization Science 27(1), 2016, pp. 207–221.

[19] Hong, L., and S.E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers", Proceedings of the NAS 101(46), 2004, pp. 16385–16389.

[20] Jeppesen, L.B., and K.R. Lakhani, "Marginality and problem-solving effectiveness in broadcast search", Organization Science 21(5), 2010, pp. 1016–1033.

[21] Jones, K.S., "A statistical interpretation of term specificity and its application in retrieval", Journal of Documentation 28(1), 1972, pp. 11–21.
[22] Kakatkar, C., J.K. de Groote, J. Fueller, and M. Spann, "The DNA of Winning Ideas: A Network Perspective of Success in New Product Development", Academy of Management Proceedings, 2018.
[23] Kaplan, C.A., and H.A. Simon, "In search of insight", Cognitive Psychology 22(3), 1990, pp. 374–419.
[24] Katila, R., and G. Ahuja, "Something old, something new: A longitudinal study of search behavior and new product introduction", Academy of Management Journal 45(6), 2002, pp. 1183–1194.
[25] Kauffman, S., and S. Levin, "Towards a General Theory of Adaptive Walks on Rugged Landscapes", Journal of Theoretical Biology 128(1), 1987, pp. 11–45.
[26] Kauffman, S., J. Lobo, and W.G. MacReady, "Optimal search on a technology landscape", Journal of Economic Behavior and Organization 43(2), 2000, pp. 141–166.
[27] Kim, T.S., and S.Y. Sohn, "Machine-learning-based deep semantic analysis approach for forecasting new technology convergence", Technological Forecasting and Social Change 157, 2020.
[28] Köhl, A., S. Fuger, M. Lang, J. Füller, and M. Stuchtey, "How Text Mining Algorithms for Crowdsourcing Can Help Us to Identify Today's Pressing Societal Issues", Proceedings of the 52nd HICSS, 2019, pp. 689–698.
[29] Kornish, L.J., and K.T. Ulrich, "Opportunity spaces in innovation: Empirical analysis of large samples of ideas", Management Science 57(1), 2011, pp. 107–128.
[30] Kusner, M.J., Y. Sun, N.I. Kolkin, and K.Q. Weinberger, "From word embeddings to document distances", 32nd ICML, (2015), 957–966.
[31] Le, Q., and T. Mikolov, "Distributed representations of sentences and documents", 31st International Conference on Machine Learning, (2014), 1188–1196.
[32] Lee, C., D. Jeon, J.M. Ahn, and O. Kwon, "Navigating a product landscape for technology opportunity analysis: A word2vec approach using an integrated patent-product database", Technovation 96–97, 2020.
[33] Lopez-Vega, H., F. Tell, and W. Vanhaverbeke, "Where and how to search? Search paths in open innovation", Research Policy 45(1), 2016, pp. 125–136.
[34] Majchrzak, A., and A. Malhotra, "Effect of knowledge-sharing trajectories on innovative outcomes in temporary online crowds", Information Systems Research 27(4), 2016, pp. 685–703.
[35] Merz, A.B., "Mechanisms to select ideas in crowdsourced innovation contests - A systematic literature review and research agenda", Proceedings of the 26th ECIS, (2018).
[36] Mikolov, T., K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space", 1st ICLR - Workshop Track Proceedings, (2013).
[37] Montello, D.R., "Scale and multiple psychologies of space", Lecture Notes in Computer Science, (1993).
[38] Murhaf, F., A. Kutuzov, S. Oepen, and E. Velldal, "Word vectors, reuse, and replicability: Towards a community repository of large-text resources", Proceedings of the 21st NODALIDA, (2017), 271–276.
[39] Murtagh, F., "Multidimensional clustering algorithms", In Compstat Lectures. Physika Verlag, Vienna, 1985.
[40] Newell, A., and H.A. Simon, Human problem solving, Prentice-Hall, 1972.
[41] Pennington, J., R. Socher, and C.D. Manning, "GloVe: Global vectors for word representation", Conference on Empirical Methods in Natural Language Processing, (2014), 1532–1543.
[42] Rhyn, M., and I. Blohm, "Combining collective and artificial intelligence: Towards a design theory for decision support in crowdsourcing", Proceedings of the 25th ECIS, (2017), pp. 2656–2666.
[43] Rosch, E., "Principles of Categorization", Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence, 2013, pp. 312–322.
[44] Schilling, M.A., and E. Green, "Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences", Research Policy 40(10), 2011, pp. 1321–1331.
[45] Schmidt, C.W., "Improving a tf-idf weighted document vector embedding", 2019.
[46] Schunn, C.D., and D. Klahr, "A 4-Space Model of Scientific Discovery", Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society, 1995.
[47] Simon, H.A., The Sciences of the Artificial, 1996.
[48] Stell, J.G., and M.F. Worboys, "Generalizing graphs using amalgamation and selection", Lecture Notes in Computer Science 1651, 1999, pp. 19–32.
[49] Sweller, J., "Cognitive load during problem solving: Effects on learning", Cognitive Science, 1988.
[50] Terwiesch, C., and Y. Xu, "Innovation contests, open innovation, and multiagent problem solving", Management Science 54(9), 2008, pp. 1529–1543.
[51] Toubia, O., and O. Netzer, "Idea generation, creativity, and prototypicality", Marketing Science 36(1), 2017, pp. 1–20.
[52] Towne, W. Ben, and J.D. Herbsleb, "Design Considerations for Online Deliberation Systems", Journal of Information Technology & Politics 9(1), 2012, pp. 97–115.
[53] Towne, W. Ben, C.P. Rosé, and J.D. Herbsleb, "Measuring Similarity Similarly : LDA and Human Perception", ACM Transactions on Intelligent Systems and Technology 8(1), 2016.
[54] Tversky, A., "Features of Similarity", Psychological Review 84(4), 1977, pp. 327–352.
[55] Ulrich, K.T., "Computation and Pre-Parametric Design", MIT Artificial Intelligence Laboratory, 1988.
[56] Zhang, C., Y.P. Kwon, J. Kramer, E. Kim, and A.M. Agogino, "Concept Clustering in Design Teams: A Comparison of Human and Machine Clustering", Journal of Mechanical Design, Transactions of the ASME 139(11), 2017, pp. 1–9.