Promoting Scientific Creativity by Utilising Web-based Research Objects

Project acronym: Dr Inventor

Deliverable No. 2.4
Final version of report for Computational Scientific Creativity

Grant agreement no: 611383

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th>Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>PU</td>
<td>Public</td>
<td></td>
</tr>
<tr>
<td>PP</td>
<td>Restricted to other programme participants (including the Commission Services)</td>
<td>X</td>
</tr>
<tr>
<td>RE</td>
<td>Restricted to a group specified by the consortium (including the Commission Services)</td>
<td></td>
</tr>
<tr>
<td>CO</td>
<td>Confidential, only for members of the consortium (including the Commission Services)</td>
<td></td>
</tr>
</tbody>
</table>
ABSTRACT:
This deliverable is the final version of the report for computational scientific creativity. This report covers task “T2.5: Theoretical modelling of computational scientific creativity”. This report also aims to overview and summarise the main findings of the project in relation to computational creativity. The analogy-based model of creative scientific think worked quite well on the lexically sourced data, forming reasonable analogies. Analogy quality for this creative task surprisingly related to the inferences more than the analogical comparison per se. Achievements include significant progress towards the publication of a paper based on a Dr Inventor created idea, as well as achieving a positive score on the Creativity Support Index (CSI). Challenges encountered and some future directions are outlined.

KEYWORD LIST:
Conceptual blend, mapping, counterpart projection, analogy, inference, blended space, scientific creativity, evaluation.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7-ICT-2013.8.1) under grant agreement nº 611383.
The author is solely responsible for its content, it does not represent the opinion of the European Community and the Community is not responsible for any use that might be made of data appearing therein.

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Status</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>19/12/2016</td>
<td>Draft</td>
<td>D. O’Donoghue</td>
</tr>
<tr>
<td>2.0</td>
<td>7/3/2017</td>
<td>Draft</td>
<td>D. O’Donoghue</td>
</tr>
<tr>
<td>2.1</td>
<td>13/3/2017</td>
<td>Draft</td>
<td>D. O’Donoghue</td>
</tr>
<tr>
<td>2.2</td>
<td>27/3/2017</td>
<td>Draft</td>
<td>D. O’Donoghue, Yalemisew Abgaz</td>
</tr>
<tr>
<td>2.3</td>
<td>20/4/2017</td>
<td>Draft</td>
<td>D. O’Donoghue</td>
</tr>
<tr>
<td>2.4</td>
<td>26/4/2017</td>
<td>Final</td>
<td>D. O’Donoghue</td>
</tr>
</tbody>
</table>

List of contributors

− Diarmuid P. O’Donoghue, Maynooth University, Ireland.
− Yalemisew Abgaz, Maynooth University, Ireland.
− Donny Hurley, Maynooth University, Ireland.

List of Internal Reviewer’s

− Horacio Saggion, University Pompeu Fabra, Spain
− Francesco Ronzano, University Pompeu Fabra, Spain


1 Executive Summary

This deliverable describes the insights that have been gained into modelling one mode of creative scientific reasoning, that of analogy-driven creativity. We outline some key findings related to the implementation and development of a type of creativity support tool, which we call a Creativity Enhancement Tool (CET). We also outline some key features of the model that we believe, led to Dr Inventor identifying high-quality analogical comparisons and thereby, better serving to enhance its users’ creativity. We review the user engagement activities and particularly the user evaluations of its computationally created artefacts. We also briefly look at some wider issues related to computational modelling of creative scientific reasoning, focusing on some advantages and disadvantages arising from the use of a CET such as Dr Inventor.

One objective mentioned by the senior scientific officer at the start of the initial meeting related to the Creative Achievement Questionnaire (CAQ), is asserting that one level of creative scientific achievement was a contribution to a published scientific paper. While raising this as a possible objective for this project was greeted with some surprise at the time, this project can report significant progress towards this goal. Partner BU adopted a research hypothesis created by Dr Inventor, developing it into a full research project and now have a paper (almost) ready to submit to a journal. Details on this work in progress are included in the sister document “D8.6 Final version of Evaluation Report”, with a brief summary contained in this deliverable.

1.1 Summary of the Novelty

The majority of this report concerns novel information and findings. We examined a list of publications deemed to be creative by experts. Results indicate that creative publications tend to attract higher levels of citations than comparable publications.

We discuss the impact that some of the lexio-semantic features of Dr Inventor had upon the ROS-graphs that were generate, especially the impact of ROS topology. We outline how graphs topologies grow super-exponentially with the number of edges (verbs) in a graph, assessing its impact on mapping larger documents. We describe how some detailed computational metrics for analogy can support the identification of creative comparisons. Despite our expectations we discovered that expert ratings for creative comparisons appear not to rely on the strength of the analogy(mapping) but more so on the novelty of the resulting inferences. We outline some of the feedback obtained on the SIGGRAPH analogies.

We look at some independently devised metrics for attributing creativity to a tool or process. We calculated Dr Inventor’s Creativity Support Index (CSI) (Cherry & Latulipe, 2014) score of 70.266 (out of 100), comparing quite favourably to other support tools. We also briefly outline the Creative Achievement Questionnaire (CAQ) showing how the “Flowers Paper” currently being completed by the BU partner (as described in other deliverables) based on an original idea by Dr Inventor will (shortly we expect) result in a rating of 4 on a 7-point scale for creative scientific discovery. The top point on this scale (7) indicates that the publication has been cited by other authors, which also becomes a possibility should the paper be accepted for publication.

We conclude by looking at some wider avenues for future progress.
2 Introduction

2.1 Purpose of this document

The purpose of this document is to summarise the main outcomes of task "T2.5 Theoretical modelling of computational scientific creativity: PM1 – PM36". It looks at one specific mode of creative scientific thought, that of creative analogical reasoning. We assess the model, how well it supports creative scientific users and what this tells us about creative scientific thinking and about creativity.

2.2 Relationship to Other Deliverables

In this deliverable we look back at select contents from previous deliverables, highlighting the most relevant contributions as well as identifying potential areas for future progress.

![Figure 1 Relationship between Deliverables](image)

D2.2 - Report for ROS
D2.3 - Initial version of report for computational scientific creativity
D3.1 - Initial version ROS implementation with report
D3.2 - Final version ROS implementation with report
D3.3 - Implementation of analogical reasoning, transformation and conceptual blending with report
D8.3 - Final version of evaluation methodology report
D8.4 - Final version of benchmark datasets

This project hoped to achieve a number of specific objectives, but this was somewhat hampered by one specific problem. Encouraging actual users to use the system and thereby gather user feedback was a significantly more complex task than initially envisaged. It was hoped that two significant user engagement activities (29th International Conference on Computer Animation and Social Agents - CASA workshop and especially the booth a SIGGRAPH-2016) would foster reasonable levels of user engagement. This was followed by a persistent email campaign and social media engagements. Probably, the most effective means of gathering user engagement activities has been the exploitation of personal contacts, especially by BU and UB. However, gathering a large base of regular users has proved to be exceptionally challenging. (See Deliverable “D8.6 Final version of evaluation report”). Most surprising was the fact that much effort that went into developing the user experience and streamlining the front end, improving back-end performance and analogy quality appear to have
contributed little to gathering a large user base. It was expected that even “idle curiosity” would gather a larger number of passing users, but this does not appear to have happened.

Upon reflection, one constant thread of our user engagement activities has been the description of Dr Inventor as a “creativity tool”. It has previously been noted that people can have a bias against computer generated artefacts, such as music (Moffat and Kelly, 2006). However, the users that did use the system seemed to value it, judging by such simple metrics as average session time (almost 10 minutes) and various feedback and evaluation questions outlined in this report.
3 Scientific Creativity

In this section we outline some wider issues related to computational creativity, arising directly from the Dr Inventor project. The Dr Inventor tool aims to support and encourage creative scientific thinking through the use of a computationally created analogies between research publications from the area of computer graphics. Such analogies and their corresponding inferences form an hypothesis for presentation and evaluation by a domain expert – with experts in the area of computer graphics being the focus for this project.

Analogy (and metaphor) research has recently suggested that metaphorically based comparisons can recruit activation from the somatosensory cortex when processing sentences containing textural metaphors (Lacey et al., 2012). So it seems there is a deep neurological bases to the interpretation of analogies/metaphors that is different from literal sentences. A more recent study on child development on analogical processing tasks indicates that improvements in analogical comprehension “is driven largely by improvements in the ability to selectively retrieve task-relevant semantic relationships” (Whitaker et al., 2017). We may well think of the Dr Inventor project as identifying similarities between publications that is based on identifying similar semantic relationships between documents.

Before looking at Dr Inventor system we briefly review some related systems from the topic of computational creativity.

3.1 Background and Creative Systems

Recent years have seen an upsurge in interest for areas like artificial intelligence and even the sub-discipline of computational creativity. This document is rooted in analogical reasoning and its role in scientific creativity and progress.

The established approach of Literature Based Discovery (LBD) (Bruza & Marc, 2008) is also arguably a creative undertaking, whose ABC model aims to identify specific knowledge (B) that connects two distinct bodies of literature (A and C). Cross-Context Bisociation Explorer - CrossBee (Juršič, Bojan, Tanja, & Lavrač, 2012) adopts an LBD-like approach and identifies semantically founded terms that act a connections (or bridges) between the words occurring in two documents in distinct areas of research. CrossBee uses term ranking based on the voting of an ensemble of heuristics, using a gold standard of LBD documents to identify the most useful heuristics from their initial list of over 40 heuristics. CrossBee employs a Bag-of-Words representation which is also used by the MAC/FAC (Forbus, Gentner and Law, 1995) analogy retrieval model that was developed to accompany the SME (Falkenhainer, Forbus, and Gentner, 1989) analogy model. While CrossBee looks for (isolated) connecting terms Dr Inventor also searches for connecting terms – but focuses on larger systematic collections of connecting terms. Additionally, Dr Inventor employs graph matching algorithm to ensure the identified similarities contribute to analogical comparisons. This graph matching sub-process allows Dr Inventor to incorporate some comparisons terms (akin to B terms) with no identified pre-existing similarity between them.

One of the most impressive achievements by an (arguably) computationally creative system has been that of Knowledge Integration Toolkit (KnIT). KnIT (Spangler, et al., 2014) aims to predict scientific discoveries by analyzing past literature. It extracts and collects information from multiple publications, looking for literal similarities that focus on central topics. KnIT proposed a novel and testable hypothesis related to a tumour–suppressing protein called “p53”.

“...identify new protein kinases that phosphorylate the protein tumor suppressor p53.
Retrospective analysis demonstrates the accuracy of this approach and ongoing laboratory..."
experiments suggest that kinases identified by our system may indeed phosphorylate p53. These results establish proof of principle for automated hypothesis generation and discovery based on text mining of the scientific literature.” (Spangler, et al., 2014)

We note that this hypothesis is quite fact based, particularly in comparison to the more high-level hypotheses created by Dr Inventor. While Dr Inventor and KnIT focus on scientific literature, only Dr Inventor explores non-literal similarities between publications. However, KnIT’s creation of a novel and testable hypothesis is an aim that Dr Inventor wishes to emulate. KnIT and Dr Inventor seem to share a similar philosophy in that the computational tool acts in a co-creative manner to perform much of the laborious tasks involved in reading large volumes of scientific literature and from this, to present a novel hypothesis for further exploration by an experienced research scientist.

One of the most high-profile systems for cognitive computing is IBM Watson (High, 2012). This cognitive computing system incorporates a deep parsing of documents written in natural language. Watson processes large amounts of information gleaned from documents and stores derived knowledge separately. Watson has been deployed on a culinary application, combining knowledge of cooking recipes with some bio-chemical understanding, to enable a forays into culinary creativity as IBM Chef Watson (High, 2012). This system has created novel and appealing recipes, some of which have been warmly accepted. While both Dr Inventor and IBM Watson adopt broadly similar approaches by combining deep parsing with semantically derived structures, Dr Inventor’s focus on analogical comparisons (and conceptual blends) does not appear to have a parallel in IBM Watson. (Goel, et al., 2015) discuss how students used Watson first as an aid to co-creativity and subsequently as a means of actively enhancing co-creativity. These projects used Watson within a question-answering context, involving larger amount of user interaction than envisaged by Dr Inventor. Like Dr Inventor, Goel’s 6 distinct applications focused on the specific problem domain of bio-inspired design, whereas Dr Inventor focused on computer graphics.

Dr Inventor differs from systems like IBM Watson and KnIT by focusing on analogical similarities between information systems. Gentner (1983) distinguishes four categories of similarity (Table 1), highlighting a distinction between surface features and the deep structure of information. Dr Inventor searches for non-obvious analogies between publications with few obvious surface similarities (such as involving few similar objects) while involving similar structures.

Some story telling creative system with some relevance to Dr Inventor, particularly Riu (Zhu and Ontañón, 2014) system as it uses analogy-based story creation. However these system do not traditionally operate on the raw text of known stories, but instead use annotated representations. Riu also makes use of pre-authored content in the form of memories and story graphs. Thus most of the challenges addressed by Dr Inventor are avoided by Riu.

COINVENT (Schorlemmer, et al., 2014) is another concept invention system which attempts to build a formal model of conceptual blending by drawing various interdisciplinary research results. COINVENT is aimed at gaining a deep understanding of conceptual blending and developing a formal method for building a generic creative computational system. COINVENT uses mathematics and music as a working domain, while Dr Inventor focuses on analogical comparisons (but not literal similarities) between computer graphics publications.

KnIT and other systems as well as the Dr Inventor project point to a promising future for the newly proposed Grand Challenge for artificial intelligence – “to develop AI system to make major scientific discoveries in biomedical sciences that worth Nobel Prize” (Kitano, 2016). While some work has explored literal similarities between research publications, we believe that Dr Inventor is the only system exploring non-literal similarities for creative effect. Some recent achievements of the Dr Inventor project, particularly the “flower paper” discussed in Section 6.1 show the potential of our approach.
3.2 Analogy and Creativity

Dr Inventor is a system that aims to both simulate one mode of creative scientific thought while simultaneously generating outputs designed that invigorate a user’s creative scientific reasoning. As such it is an instance of bisociative reasoning (Koestler, 1964), whereby novel analogical comparisons bring fresh ideas to bear on some problem.

![An internet meme from early 2017, based on a simple (password: toothbrush) analogy](image)

Like all analogies, this comparison highlights specific aspects of the resemblance that were previously (somewhat) overlooked. Finding and exploiting such analogical comparisons using computational means is the focus of this report.

Creative ideas are themselves often described through metaphors – with those metaphors impacting our value judgements of idea quality and the inventor’s genius (Elmore and Luna-Lucero, 2016). New and potentially creative ideas can be seen through the “light-bulb” metaphor involving a sudden switching of the idea, while the alternate “seed” metaphor focuses more on the gradual progression of the new idea.

Analogical thinking can be seen within a slightly wider context of similarity. Gentner (1983) sees analogy as a particular form of similarity involving two analogs that share few similar object (and attributes) while sharing some relational structure (Table 1). Thus analogy involves non-obvious similarity with Gentner's structure mapping theory (SMT) initiating decades of research to elucidate their particular form of reasoning. Later in this report we will show how metrics like relational similarity (relSim) and conceptual similarity (conSim) can be used to detect different forms of similarity, such as analogy (though it may be best to think of relational and conceptual similarity lying on different ends of a continuum).

<table>
<thead>
<tr>
<th>Similar attributes and objects</th>
<th>Similar relational structure</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literal similarity</td>
<td>Many</td>
<td>Many</td>
</tr>
<tr>
<td>Surface similarity</td>
<td>Many</td>
<td>Few</td>
</tr>
<tr>
<td>Analogy</td>
<td>Few</td>
<td>Many</td>
</tr>
<tr>
<td>Dissimilar or Anomaly</td>
<td>Few</td>
<td>Few</td>
</tr>
</tbody>
</table>

It has been noted that scientists use analogies in different ways, some using lots of little analogies (eg Robert Boyle late 1600’s) often putting forth several analogies for each principle he wished to prove (Gentner and Jeziorski, 1993). In contrast, others employed a small number of deep and expansive analogies (eg Keppler early 1600’s). Of course individual publications will generally not encompass all analogies used by an individual author. Gentner and Jeziorski (1993) also note that scientific progress must treat analogical inferences as conjectures – and not as already proven facts (as the alchemists). Our focus is on developing a co-creativity (enhancement) tool, retaining the "human in the loop" and allowing proper scientific progress as all suggestions will be evaluated by subject experts.
An informal reading of a sample of documents from Dr Inventor’s SIGGRAPH corpus shows that most papers focus on a single problem offering one main approach to solving that problem. This should in principle allow identification of potentially fruitful analogies form by (say) the approach developed in one paper with the problem described in another.

3.3 Creative Papers Assessment

Before looking at the Dr Inventor tool, we look at an objective analysis of creative papers showing that creative papers appear to differ quantitatively from “normal” papers in a few important aspects. An open survey was spread among the consortium partners and their personal connections, collecting bibliographic details on two types of creative papers:

1) Random: This collection serves as our baseline for comparison against two categories of creative paper. The list of papers available from DBLP was downloaded collection and 50 papers were randomly selected, using an online pseudo-random number generator to generate the index used to select these papers. DBLP predominantly covers papers from the discipline of computer science, including many journals and conference publications. In order to allow comparison between the random DBLP papers and the creative papers, we removed from the creative papers from beyond the computer science (CS) discipline, leaving a collection of creative CS papers.

2) Own: These are papers whose bibliographic details were collected during the course of a data gathering activity. These papers were actually written by the respondents, but are ones they consider to be their most creative – but do not necessarily appear in highly ranked publications venues.

3) Known: These are papers submitted by the same respondents as 2 above, but are papers that respondents had read and that they deemed to be particularly creative. But these papers were not written directly by the respondents themselves. As we expected the respondents to have read more papers than they had written, we expected this collection to contain the most creative papers.

Work on the ROS Assessment Tool (ROSAT) (originally mentioned in D3.1) was completed and used to compare the two categories of creative papers against the random DBLP papers.

Citations: The boxplot in Figure 3 (a) depicts the number of citations received for the randomly selected DBLP papers with the respondents (own) most creative papers. As can be seen the respondents own creative papers occasionally receive more citations. However, a Wilcoxon Sum Rank test revealed no statistically significant difference between these collections.

Figure 3 (b) compares all three categories of paper, with this boxplot showing that the creative papers sourced from the wider literature received far higher numbers of citations than previous two categories. This was supported by a Wilcoxon sun rank test showing the difference between the Own and Known categories to be significant and reliable (W = 217.5, p <0.005). While perhaps not a very surprising result, this does lend support to the importance of creativity in scientific publication – at least as measured within the discipline of computer science. This we take a further support for the mission underlying the Dr Inventor project.
Year of Publication: The number of citations gathered by a publication can be heavily impacted by its age. So we followed the citation analysis by examining the age of documents in each of the collections. The following boxplot shows the distribution of ages, showing that the Known category includes some of the oldest publications. The Random and Own categories are quite compatible, with the Random category being somewhat older as expected. However, the previous citation comparison appears to be a sufficiently robust finding, as the Known creative papers were not significantly older. Indeed, their slightly increased longevity is a further indication of the long term contributions made by creative papers of the research community.
**H-index:** The following boxplot shows the H-index for these collections, with the H-index of the publication venue coming from (Journal or Conference) by SJR - SCIMAGO Journal Rank which includes rankings for many of the major conferences - especially important for the computer science discipline. The H-index data shows that the Known creative generally originates in highest ranked publications, with the Own creative collection being next followed by the Random papers. This is the pattern that was expected, showing that the increasing creativity generally corresponds to publication in better quality venues.
**Field Rating:** Microsoft Academic offers a standard means of comparing publication venues, with its “Field rating” that encompasses both Journals and Conferences in a single metric. This metrics was selected as the discipline of computer science makes greater use of conferences as reputable publication venues than many other disciplines. Indeed, evidence suggests that the better conferences rank on a par with mid-ranking journals (Freyne et al, 2013).

Unfortunately we did not find sufficient data for each of the categories so no data or conclusions are included in this deliverable.

**Impact Factor:** The Impact factors of the publication venues were also examined for the three categories of paper. The pattern of results here was similar to that found with the H-index, with the exception that the Random and Own categories appeared to overlap to a far greater degree. However, the impact factor for the creative publication venues was the highest.
The H-index and impact factor data may suggest different interpretations. Firstly, it may suggest that the authors of creative papers were aware of their papers potential and therefore deliberately chose better journals for that publication. An alternative explanation is that any paper published in a more highly ranked journal stands an increased chance to grabbing the imagination of the community, resulting in the eventual accretion of larger numbers of citations. We argue in favour of the former interpretation and against the latter, as the results from the “Own” category appear to support the hypothesis that authors are aware of the creativity of their own publications.

**Institution Rank:** A comparison of the institution ranking for the three categories of paper shows little difference between the institutions. This result was somewhat surprising as it was expected that the known creative papers would be associated with better ranked institutions – but this does not appear to be supported by our evidence. Additionally, the Own and Known collections appear to arise in surprisingly similarly ranked institutions. This might be yet more evidence of the unpredictable nature of creativity – that it does not necessarily arise when and where one would expect.
From this analysis we see that one of the most significant differences to normal publications concerns the accrual of citations. While creativity may not be the only contributor to citations, our analysis shows creativity has a marked impact on a discipline as measured by increased citation counts. This both highlights the importance of scientific creativity and serves as further motivation for the Dr Inventor project.

1. Creative SIGGRAPH Publications

Among the gathered responses for creative papers were 8 SIGGRAPH papers - noting that a few of these papers appear as both “Transactions of Graphics (ToG)” and “Proceedings of ACM SIGGRAPH” because all SIGGRAPH publications are subsequently published in ToG. The respondents’ information included their attributed Level of Creativity for each paper, on a scale between 1 and 5 with 5 indicating the maximum level of creativity.

NOTE: at the time of writing, none of these papers were contained in the Dr Inventor corpus for one of the following reasons. 1) The date was outside the Dr Inventor range 2002-2016, 2) the file was too large for the PDF text extractor, 3) SIGGRAPH ASIA was not part of the corpus. However these creative papers appear to (generally) have large numbers of citations, as can be seen from the right-most column in the above table.

This objective analysis of papers that were suggested as being creative, highlights the importance of creativity within scientific publications. We see the creativity of a publication as often resulting in large number of citations. Thus we argue that a tool to enhance scientific creativity could (arguably) result in publications that attract larger numbers of citations.

This deliverable focuses on developing and evaluating a tool to enhance scientific creativity, one built on a partial model of analogical reasoning. But before examining Dr Inventor in detail we briefly look at general frameworks for computational creativity, examining the case for Dr Inventor to be seen as a model of computational creativity.
<table>
<thead>
<tr>
<th>Title</th>
<th>Citations</th>
</tr>
</thead>
</table>

### 3.4 Creativity when Reviewing Papers

We also sought expert opinions on the importance of creativity, when reviewing papers for acceptance in journals and conferences. Expert feedback was sought on for the following three questions:

1. Creativity is important when reviewing papers
2. I can assess the level of creativity in a paper
3. I can compare the levels of creativity between two papers
Interestingly, creativity seemed to be rated more highly by more experienced researchers (Professors and Senior Lecturer) who gave an average rating of 4.7 (of 5) for Importance (Q1).

Respondents also offered their opinions on the Scientific Discourse Ontology (SDO) section of a paper where the creativity becomes most apparent (Background, Challenge, Approach, Outcome or Future Work). This question focused on the two central qualities of creativity, namely novelty and quality (or usefulness).

The challenge and the approach seem to be the sections that are most strongly associated with creativity. Not surprisingly, the quality seems to be best evaluated in the “outcome” section of a paper.

Jordanous (2016) identifies 14 independent qualities of general creativity, however some of these are less central to scientific creativity than others (e.g. emotion and self-expression). Further feedback was sought on expert perceptions of the importance of these factors to scientific creativity. Some subsequent evaluation activities focused on three most relevant qualities, highlighted in green below.
3.5 Literature Search for Creativity

Finally, we asked respondents about their current practices in searching for creativity. Three questions were presented, the firsts asking if they believed there were Publications in other disciplines that were relevant, but they had not found them. Secondly, they were asked if they deliberately seek papers from other disciplines. Finally, they were asked if they were provided with a novel creativity support search tool, would they prefer inaccuracy to miss potentially useful document. Responses are summarized in Figure 7 below.

![Figure 7: Responses to three questions related to a creativity support tool](image)

3.6 Frameworks for Creativity

Computational Creativity is frequently described using Boden's (1994) three-level framework for creativity (1) Improbable/Combinational, (2) Exploration and (3) Transformation, noting the underlying analogy (metaphor) of spatial exploration that unites these three levels. Firstly we note that each
analogue comparison (mapping) identifies the largest common subgraph between two papers (as represented by their ROS graphs), tempered by a preference for mapping semantically similar nodes.

The number of potential mappings between two ROS graphs scales exponentially with the number of nodes in each graph. The number of potential "combinations" (i.e. analogies) scales linearly with corpus size for any given target. Next, we look at combinational creativity arising from the realization of analogical inferences. Dr Inventor identified 26,072 distinct concepts and 7,315 distinct relations (discounting repeated instances of each) from the SIGGRAPH corpus. This allows creation of $4.97 \times 10^{12}$ distinct subj-verb-obj combinations. Furthermore, exploring clusters of just 5 inferences would yield $3.03 \times 10^{63}$ possible inferences, while the presence of co-references increases the space of possible inferences yet further. As comparison the number of atoms in the observable universe is generally taken to be $4 \times 10^{79}$. Thus, the combinatorial challenge of finding practically useful and challenging comparison is significant. However, Boden's combinatorial perspective does not highlight some aspects of this challenge.

We also contrast Dr Inventor's creativity with Howard Gardner's (1993) Big-C and little-c; Big-C being the historically eminent creativity of Newton, Einstein and others, while little-c is the more mundane everyday creativity. Of greater relevance to Dr Inventor is the more recent but related level Professional creativity (Pro-c) (Kaufman & Beghetto, 2009) "developmental and effortful progression beyond little-c that represents professional-level expertise" which doesn't achieve the eminence associated with Big-C Creativity. We argue that Dr Inventor aims to support users' creative objectives, where those users are practising researchers between the level of postgraduate research student to senior lecturer and Professor. In fact we have found that those who get greatest benefit from Dr Inventor have been senior researchers and Professors in the target discipline of Dr Inventor's main corpus of SIGGRAPH papers (2002-2016). Thus we argue that Pro-c is the most appropriate framework in which to view the Dr Inventor project.

3.7 Conclusion

In summary we believe that creativity is not only of theoretical interest to scientific publication, but also seems to be associated with higher levels of citations. Additionally, the creative papers appear to continue gathering citations over a longer timeframe than other publications. We also showed how the Professional-creativity level (Pro-c) seems to be a better fit to the objective of the Dr Inventor project than many other creativity frameworks. Later in this deliverable we shall briefly re-visit this frameworks topic after a discussion on the quality of analogies and inferences.
4 Computational Model for Creative Scientific Comparisons

This deliverable focuses on the Dr Inventor model for enhancing scientific creativity. This section briefly outlines a few of the most significant contributions that were made in modelling the analogy process, with the aim of better enhancing the users’ creativity by producing better more human-like comparisons.

4.1 Single Instance of Concept Nodes

An early change made to Dr Inventor’s text processing pipeline concerned identification of co-referent terms within text - and possibly beyond (Connell and Lynott, 2014). This was crucial in supporting mappings between the meaning of text, rather than the raw text itself. Terms and their co-referents are represented uniquely in each ROS, effectively overcoming the lexical distinction between multiple instances of a term like “montage” and a co-referent like “it”. We briefly outline the importance of these co-referent terms upon graph topology (in particular).

The Dr Inventor corpus currently holds SIGGRAPH, ICCC and Patent collections, containing 793,445 distinct concept nodes. It also holds 718,214 distinct verbs, but all verbs are uniquely represented and so are not relevant to this discussion. Of the concept (noun) nodes, co-referent terms account for 117192 concept nodes, of which there were 65259 distinct subjects and 96040 distinct object nodes. While it appears that only 16.3% of terms are co-references, these co-referent nodes have a disproportionate impact on the resulting mappings (over singly referenced concepts). The following is an example of the impact of co-references identified from the Karla:Zerdia analogy showing four distinct terms from the Karla story, corresponding to 8 distinct words in this analog.

Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot a deer instead.

Co-Referent ROS terms | Analogous Source terms
--- | ---
karla_hawk/she/her | zardia/she

We note also that each of these terms is a potential concept node in their own right, but as merged as they refer to the same concepts “in the real world”. Thus, they merge neatly into Gentner’s Structure Mapping Theory (1983). The multiple terms in ROS nodes that are discussed later in this report to support causal relations are significantly different in character and the two types should not be confused.

As well as accounting for multiple terms from an analogy, co-references have an even more obvious impact on the topology of the resulting ROS graphs. The term “montage” is used in several predicates in the Figure 8 below. Crucially, this concept is uniquely represented so that all instances within this one paper are represented by the same concept node. This ROS/paper will therefore best form an analogy with another paper containing a similar structure involving multiple references to one central concept. Relational concepts (like “be” below) are each represented distinctly, and can potentially each map to different relational verbs.
We note that were also 415,963 “Unknown” nodes where either the subject or object of a relation could not be explicitly identified. Figure 9 below displays a portion of a ROS, where the blue nodes represent concepts and the green nodes represent relations. As can be seen, one of the most pressing issues was to identify the correct referent of any “unknown” nodes (7 instances), where the corresponding argument was not successful identified. While much progress was made during the course of this progress, many unknown nodes remained.

Despite the incomplete nature of some of the derived ROS graphs, the recognition of duplicate and co-referenced conceptual nodes had the single biggest impact on the quality of the analogies that were generated.
Figure 9: Fragment of two ROS for two different papers showing multiple referenced concept nodes (blue) as both subject and object. Also shown are multiple "unknown" nodes with relation nodes (green) connecting concept nodes. Two instances of the term “montage” are displayed arising from two different papers, with only one reference being found in the second (lower) paper.
The above graph details the percentage of concepts nodes in each graph that were co-reference nodes. Of course, each co-reference node might combine multiple concepts. Due to their impact on topology, we might expect these nodes to generally experience greater involvement in analogies – as the inclusion of each might contribute to the mapping between multiple relations.

### 4.2 Representing CAUSAL Relationships

During the first half of 2016 the UPF partner completed their work on explicit identification of causality and causal relationships, extracted from the text of SIGGRAPH and other documents. After a significant amount of work they managed to identify causal relations from text, using a series of rules.

> "Since kinematic constraints can usually be represented by single equations, they can be easily embedded into optimization problems for motion synthesis. However, extension to motions involving character shapes seems difficult since relationships between rigid bodies or surfaces need to be encoded. Further, these methods can not handle close interactions without any tangles. The proposed method considers the relationships among rigid body parts and is more general since it can handle motions of close interactions with / without tangles."
Some significant problems were encountered when we attempted to integrate the causal relations into our pre-existing computational system for analogical reasoning. We firstly highlight the incompatibility between the causal relationship as described above and the ROS graphs as specified by the Maynooth team – and as used throughout the analogy retrieval, mapping and inference processes. The structure of ROS graphs was specified in Deliverable “D2.2 Report for ROS” (month 6) as a specialised form of attributed relational graph. This design aimed specifically to support the Structure Mapping Theory of Gentner (1983). ROS graphs are formed from noun-concepts interconnected by verb-based relationships, as derived from the dependency-tables produced by the dependency parser.

Supporting these causal relationships caused the text-processing team to move beyond the previous representation of their parsers output (the API’s that were fundamental to the ROS generation and implicitly to all subsequent parts of the creativity model was even deprecated at this stage). Instead their new parser output large composite nodes containing multiple terms that were not tightly aligned with the previously identified Subjects and Objects. This change was brought about because the process of identifying causal relations typically operated on phrasal structures rather than isolated terms (such as nouns and verbs). So our previous process for generating ROS became ineffective. Incorporating this new causal information into Dr Inventor’s creativity engine would have required a complete re-write of our mapping algorithm – which was beyond the scope of this project.

We illustrate a fundamental issue with the identification, representation and use of causal relationships in the Dr Inventor model.
**Cause:** 'relationships between rigid bodies or surfaces need to be encoded'

**Effect:** 'extension to motions involving character shapes seems difficult'

The crux of this issue involves not so much the identification of the causal relation itself but rather, identification of the agent and the effect of that causality.

The issue of including causal relations may well involve far more detailed reasoning than is available to a simple, domain independent, model of analogical thinking. (Holyoak & Lee, 2017) “Psychological evidence suggests that analogical inference often involves constructing and then running a causal model”. They also discuss how analogy models typically represent causal relations in the manner of this project, adding connection in a “predicate –calculus-style description” noting that this is clearly an oversimplification. They forward the idea of using more complete causal models, employing Bayesian reasoning with these models.

### 4.3 Ontological Tagging

Sentence for documents in the SIGGRAPH corpus have been tagged with their SDO ontological category. These ontological categories correspond to overlapping subsets of the full ROS allowing Dr Inventor to explore analogies between different subsets of the full papers. Additionally, Dr Inventor retained access to the author’s original lexical categories, such as Abstract, Introduction, Background and Conclusion. Dr Inventor explored analogies formed from various combinations of lexical and ontological categories. The main analogical comparisons explored by Dr Inventor were the following:

<table>
<thead>
<tr>
<th>Target</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>Abstract</td>
</tr>
<tr>
<td>Abstract + Background</td>
<td>Abstract + Background</td>
</tr>
<tr>
<td>Abstract + Approach</td>
<td>Abstract + Challenge</td>
</tr>
<tr>
<td>Abstract + Approach</td>
<td>Abstract + Approach</td>
</tr>
<tr>
<td>Abstract + full-text</td>
<td>Abstract + full-text</td>
</tr>
</tbody>
</table>

However, the distribution of nodes between the different SDO categories was far from equal. While some categories barely contributed any nodes to an average ROS, others were overwhelmingly large (e.g. 70%). Clearly, this can be attributed to the discipline of computer science and the graphics sub-discipline, but can also be partly attributed to the SIGGRAPH conference series. We might well expect a very different distribution to arise from non-computer science publications, especially more experimental ones.

Due to difficulties gathering user evaluations, the majority of this deliverable focuses on the Abstract to Abstract comparisons. However, while NCCA from the BU partner were preparing their “flower” paper from a research hypothesis created by Dr Inventor, they were given access to several analogies between the full-text of different papers. As shown in Figure 10 the analogies between full text of two papers is also available on the live system. During development of the “flower” papers discussed later, full-text mapping proved to be particularly useful in understanding the details of some analogies.
4.4 Theoretical Perspective on the Central Mapping Process

The central process to our creativity model concerns analogical mapping, conforming to the structural similarity and systematicity principles of (Gentner, 1983). That is, an analogy can be best seen as arising from structurally similar structures used in two semantically dis-similar graphs. This is the key insight into seeing one concept can be seen as bearing some similarity with an apparently dissimilar concept.

While initial investigations explored traditional “incremental” algorithms (such as used in O’Donoghue and Keane, 2012), development quickly switched to the VF2 (Cordella et al, 2004) algorithm due to its noted quality and efficiency, as well as its ability to combine semantics with topology into the process of identifying common subgraphs. An assessment performed by mapping over 1000 graphs to themselves, showed that our VF2-based algorithm always generated the maximal mapping for such tests. We note the "tailorability concern" raised by Gentner and Forbus (2012) raising concerns that analogy models may become overly focused on a small range of problems, but we believe Dr Inventor does not suffer from such a limitation and could in adopt a different mapping model such as SME (Falkenhainer, Forbus & Gentner, 1989; Forbus, Ferguson Gentner, 1994) to generate qualitatively comparable results.

The following table summarises the run-time performance of our VF2 based mapping algorithm, running on our local server. Even for large graphs the run time is typically around 30 seconds for each pair of graphs – taken from a large sample of SIGGRAPH 2002-2015 graphs.

To compare the similarity between terms, initially we used the WU & Palmer (WUP) (Zhibiao & Martha, 1994) similarity metrics. After a continuous comparison of other term to term similarity measures, we found that Lin (Lin, 1998) similarity measure is more suitable to generating creative comparisons.

<table>
<thead>
<tr>
<th>Analogy Type (using SDO)</th>
<th>Size of graph</th>
<th>Mapping Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract &lt;&gt; Abstract</td>
<td>Min = 3</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Mean = 12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Max = 42</td>
<td></td>
</tr>
<tr>
<td>Abstract + Background &lt;&gt; Abstract + Background</td>
<td>Min = 5</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Figure 11: Full text analogies were made available for each analogy.
<table>
<thead>
<tr>
<th>Mapping Type</th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract + Approach &lt;-&gt; Abstract + Challenge</td>
<td>9</td>
<td>256</td>
<td>887</td>
<td>1.13</td>
</tr>
<tr>
<td>Abstract + Approach &lt;-&gt; Abstract + Approach</td>
<td>34</td>
<td>628</td>
<td>1101</td>
<td>19.0</td>
</tr>
<tr>
<td>Abstract + full-text &lt;-&gt; Abstract + full-text</td>
<td>36</td>
<td>851</td>
<td>1353</td>
<td>34.2</td>
</tr>
</tbody>
</table>

Semantic similarity measures between a possible mapping \( (W_s : W_t) \) used a Java application to compute these scores, but analysis showed this was often the slowest part of the mapping process. To improve performance all returned similarity measures are cached in a local database, so that any subsequent similarity between the same terms can be quickly retrieved. This modification had the biggest impact on run times, allowing larger mappings to be explored within any given timeframe.

![Figure 12](image.png)

**Figure 12** Relations involving the verb “be” for a single ROS of a full paper

Developing mappings between two full ROS graphs raises another issue about graph similarity. Figure 12 (above) shows all instances of a single relational verb “be” within a single ROS. Mappings between two graphs can easily become focused on details that are not central to the main argument of a paper.
– due in part of the exponentially increasing number of possible topologies that may be formed from some number of typical relations. We note also that this problem is exacerbated by the exponential distribution (aka Zipf’s law) describing the frequency with which terms are used.

Figure 13 Part of a patent ROS showing several instances of the verb “be” between the same arguments, along with a few other relations between those two arguments

We estimate the number of possible topologies for a graph with N edges using the following formula:

\[
\text{Number of Possible Topologies} = 2^{N(N-1)}
\]

<table>
<thead>
<tr>
<th># Nodes</th>
<th>Possible Topologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>64</td>
</tr>
<tr>
<td>4</td>
<td>4096</td>
</tr>
<tr>
<td>5</td>
<td>1048576</td>
</tr>
<tr>
<td>6</td>
<td>1073741824</td>
</tr>
<tr>
<td>7</td>
<td>4.39805E+12</td>
</tr>
<tr>
<td>8</td>
<td>7.20576E+16</td>
</tr>
<tr>
<td>9</td>
<td>4.72237E+21</td>
</tr>
</tbody>
</table>

Figure 14 Super Exponential Growth in the number of possible topologies for N edges

We note the logarithmic scale on the vertical axis plotting the exponential growth in the number of possible topologies arising from N directed edges. Given this explosive rise in the number of possible topologies we might legitimately expect that as graphs grow in size (from Abstract to the full text of a paper), the proportion of two analogous graphs that are covered by the identified common subgraph will generally tend to reduce. That is, the larger the graphs the smaller the proportion of those graphs that will participate in the analogy. Looking at this from an alternative perspective we might theorise that the probability of finding two large documents (say two books) that are highly analogous and whose Largest Common Subgraph involves the majority of edges from the two document is exceptionally unlikely. We might therefore expect that there may be some practical boundary on the length of text (ROS graphs size) that can form a viable mapping between two documents.

Of course, our graph matching process is constrained by factors beyond topology (ie semantic similarity of mapped edges), so we might think of the above analysis as a “worst case scenario”. But
the highly exponential growth in the space of possible topologies serves as a practical limit constraining the size of the expected Largest Common Subgraph (LCS) that might be identified between two analogous documents. Allied to this is the fact that similar topologies may become an increasingly powerful means of detecting plagiarism (and other forms of near-identical similarity) between documents – rooted partly in the expected degree of similarity one might generally expect to find in two randomly chosen documents of comparable size.

1. **Reliability of Analogy Metrics**

One undocumented feature of the Neo4j database related to the construction of graphs and resulted in somewhat unpredictable behaviour of our mapping process. When the same ROS graph was added to Neo4j a second time, it occasionally did so in a manner that subsequent Neo4j SPARQL queries returned items in a different order. While mapping a graph to itself always produced optimal graph-matching results, identifying maximal matching for graphs. However mapping a graph to a newly created (second) instance of that graph occasionally produced sub-optimal results due to differences in the order that data was retrieved from Neo4j.

Recent modifications were made to the ROS graph creation process and to the mapping process to address this issue. While testing is not complete we believe that we have gone a long way to resolving the issue of occasional sub-optimal mapping results. All recent attempts to re-produce the unreliable analogy metrics have (thankfully) proven unsuccessful.

![Figure 15: Two different visualisations provided for the inter-document mappings.](image)

2. **Symmetry of Analogy Scores**

We make one brief point about the reliability of analogy scores and the order in which the tenor and the vehicle. We merely wish to point out that presenting a text as either a tenor or a vehicle has an impact on the resulting similarity score. This is not an error in the metrics or the mapping process, but a deliberate design decision rooted in the theory of analogy.

Firstly we note that the analogy process is not symmetric. For example, saying that “a man is like a pig” is to say that a man lives in a sty (and is unclean). Conversely say that “a pig is like a man” is to imbibe the pig with some percepts of humanity – that the pig possesses intelligence. So using X to describe Y is generally not the same thing as using Y to describe X.

Consider the well-known creative analogy sued to describe the structure of the atom, using the analogy to the (structure of) the solar system. The prevailing theory describing the structure of the atom at this time was J.J. Thompson’s Plum Pudding model proposed in 1904. In this model the electrons correspond to the plums, which were embedded and randomly distributed through in the larger pudding. The Rutherford-Bohr analogy to the solar system used Thompsons (plum pudding) atom as the target and the solar system as the source.
If we adopt a converse perspective and swap the roles of the source and target, whereby the plum pudding model acts as the source and the solar system forms the target.

To depict this in relation to the Dr Inventor system, we use the *Karla : Zerdia* analogy as our example (in the order *Target : Source*) taken from (Gentner and Landers, 1985). We note that in the original presentation of this analogy, the inferences were already included in the description of the target domain. So this comparison forms a good test of our similarity detection mechanism. The first order of presentation to the Dr Inventor system yields a similarity score of 0.154.

![Analogical Similarity Scores](image)

However taking Karla as the source and Zerdia as the target, we get a slightly different result with an analogical similarity score of 0.19870496187836695.

![Analogical Similarity Scores](image)

Of course, we could alter our mapping system to always produce the same score regardless of which text was the source or target. But this would negatively impact the run-time of the matching process and thereby constrain the creativity that could be explored. Treating the roles in this way should result in more “human like” behaviour and the resulting comparisons should in theory, be more readily adopted by users.

3. **“Try Your Idea” - Mapping Performance Improvement**

One of the limitations of the system as described above, was that all comparisons were generated between documents that were already contained within the corpus. However, we also wished to attract users who had not published in SIGGRAPH, including those who might wish to publish in SIGGRAPH. The “Try Your Idea” (TYI) service allowed users explore analogies between a submitted
passage of text (such as the Abstract of a paper that is under preparation) and get analogous ideas from Dr Inventor. The "Try your Idea" functionality was located on the system dashboard.

During the CASA evaluation task, one user noted that this would also help her find a “template paper” to serve as a guide to writing her first journal paper. Unfortunately, subsequent attempts to contact this user did not prove successful.

Feedback from Bedfordshire University indicated that the mapping process was generally running too slowly to allow the system to identifying candidate sources “while the user waits”. To help expedite the mapping process and thereby by support faster analogies to be identified, NUIM modified the mapping process to run on multiple cores as this was identified as the main bottleneck. Crucially it was intended that this would also allow creative of better analogies between ROS, in that a greater quantity of larger ROS could be processed in any given timeframe. So the VF2 based mapping algorithm was modified to take advantage of multi-core architectures and thereby improve the Dr Inventor system.

4. Analogical Inference

As we shall see later in this deliverable, the creative impact of a comparison on a user is focused on the inferences - and not on the mapping as was initially expected. Indeed, many expert users focused much of their attention on a very small number of the more interesting (and challenging) inferences. Thus, one might think that inferences alone (even isolated inference) might drive users creativity – removing the need for the preliminary mapping/analogy process. But of course, inferences are generated by a pattern completion process which can only operate once the analogical comparison has been sufficiently detailed.

5. Inference Presentation

Inferences are presented within the context of the two analogous papers, allowing deeper understanding of the comparison and its inferences. Some key refinements made to the user interface include:

- Consistent use of colour, using orange indicating the source material and blue indicating information that originated in the target paper.
- Highlight the sentences within their original context
- Express inferences by incorporating the local lexical context of each of the SVO terms
We assess the novelty of inferences in two distinct manners. Firstly we compare inferences to the information already gathered from the entire Dr Inventor corpus. Secondly, we compare inferences against all the other inferences arising from all other analogical comparisons. This second type of analysis was undertaken to distinguish between inferences that were likely to arise from analogies within the corpus. In the next Chapter we shall refer to these two types of novelty assessment as "ObservedNovelty" and "PredictedNovelty" respectively. This novelty information was used to determine the order of presentation of the inferences contained in the user interface. This helped overcome some previous criticisms that the inference list too frequently began with creatively unproductive inferences like "We propose algorithm", putting users off further exploration of that analogy.

An n-gram approach was adopted to evaluate the inferences, with a 3-gram being sufficient to detect previously occurring instances of an inferences within a collection. This was sufficient for detecting previous instances of some highly frequent inferences like “We Propose Algorithm” (in Subject Very Object notation).

Inference generally combine one or two pieces of information from one paper with the remained of the inferred triple (S-V-O) coming from the other paper. The majority of inferences were assessed by the corresponding 2-gram models that were constructed from the two collections (SIGGRAPH corpus and the corpus of Inferences). This allowed piecemeal evaluation of inferences in terms of the bi-grams: <s> Subj, Subj Verb, Verb Obj, OBJ </s> allowing far greater coverage of the inferences.
4.5 Conclusion

Our analogically founded computational model of creative scientific reasoning was constructed and made freely available online. Several outreach activities (at CASA 2016 and SIGGRAPH 2016) were undertaken to attract an audience of users, coupled with email, Facebook and Twitter campaigns. In the next section we will see how inference novelty and many other metrics correlate with the creative impact of the analogical comparisons between scientific documents.
5 Results

In order to identify the best possible analogical comparisons that results in the greatest creative impact, we undertook a user evaluation study. Users evaluated analogies for three qualities associated with scientific creativity: Novelty, Challenge the norms of a discipline and Quality.

As stated earlier analogy is a particular form of similarity, involving similar relational structure and different objects. Much of the literature on analogy assumes that good analogies are quite rare, with people having difficulty in “coming up with” good and novel analogies. Before looking at issues related to creativity, we look at the results of Dr Inventor’s ability to assess similarity.

This chapter contains a broad spectrum of evidence that aims to give an accurate depiction of the creative achievements of the Dr Inventor system. Most of this focuses on analogies identified from the corpus of SIGGRAPH (2002-2016) papers that the analogies that were identified therein. The corpus contained 1346 documents resulting in the creation of 1146 ROS graphs, with the remaining documents not being translated into ROS graphs due to various size, parsing and other issues. We start by looking at some gross characteristics of the SIGGRAPH analogies - how common viable analogies were in terms of some simple metrics. We then look at some interesting types of similarity that can be identified by Dr Inventor. A list of computational metrics for analogies are then described, assessing semantic, structural and other factors. We the look at how some of these qualities assist in identifying the more creative analogies, as identified by users. We then look at SIGGRAPH analogies and evaluate the various forms of feedback obtained. The chapter concludes with other attempts at gathering user feedback focused on researchers on analogy and computational blending. This provides an online service for evaluating analogies between two short passages of text, giving detailed feedback including many analogy-based metrics. The other focused on the computational creativity community itself by applying Dr Inventor to a corpus of the International Conferences on Computational Creativity (ICCC) papers (2010-2016). However this effort attracted a similarly muted response from this community.

5.1 Similarity Detection – Are Analogies common?

We use the anaSim metric (described below) to identify the best source for each target, from the SIGGRAPH corpus. Part (a) of Figure 17 (below) shows the distribution of scores found for the best analogies in the corpus. Interestingly, the highest score occurred for two papers written by the same authors and related to the same topic, but published in different years. Typically around 15% of the available sources score above 0.2. The figure below shows the distribution of scores across the corpus of ~1200 documents.

![Figure 17](image_url)

*Figure 17: (a) Distribution for the best analogies found for each target document from the SIGGRAPH corpus (b) Typical scores produced for a single target paper, using all sources from the Dr Inventor SIGGRAPH corpus*
The second part of this image shows the distribution in similarity scores for the given target. We can see that only a small number of candidate source form a reasonably large mapping with the given target – despite the semantically focused collection of documents in the SIGRAPH corpus. Furthermore, we see that a small number of documents did not form any viable mapping with this target – with no paired items producing any non-zero Lin similarity score for the mapped terms.

1. **Best 500 Analogies**

Dr Inventor never expected that all its analogies would be creative – only that some or many would be creative. Thus, we focused on the better analogies that it developed. We examined the best 500 analogies produced by Dr Inventor, using an exhaustive analysis of all possible analogies from within the corpus. These best 500 analogies were selected using the *anaSim* metric (discussed later) that incorporates semantic factors (verb and noun based similarity) as well as the Jaccard coefficient to identify analogies based on a strong mapping. Most of these analogies involved similarity scores between 0.3 and 0.6. These analogies generally produced between 1 and 4 inferences.

This evaluation identified the best combination of Source:Target pairs, with 408 target ROS being used just once and 400 sources appearing just once on this list. However, some ROS appear multiple times in these best analogies. The two graphs below detail the number of times that ROS were used (1st and subsequent instances) in more than one analogy. We see that Zips law appears to apply to the re-use of sources and targets in multiple analogies. So, there appears to be a bias in our analogical mapping process that appears to favour the participation of certain ROS in analogical comparisons. Part of the explanation for this might lie in their use of common topologies or in their use of common terms that produced non-zero WordNet similarity scores.
5.2 Non-Literal Similarity and Topologically-based “Near Miss” Similarity

We briefly undertook a study to example how Dr Inventor estimates similarity between text vignettes, where the meaning of these texts are highly dependent upon the topology. Furthermore, we moderated the text similarity between the single target and the candidate source texts to highlight a specific ability of Dr Inventor and its ability to perceive similarity between texts. The table and figure below highlight the contents and structure of the target.

Table 2: Vignette and a topological view.

<table>
<thead>
<tr>
<th>Target text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom loves Mary.</td>
</tr>
<tr>
<td>Mary loves Tom.</td>
</tr>
<tr>
<td>Joe loves Mary.</td>
</tr>
</tbody>
</table>

Now consider the following four vignettes and their similarity to the above target. Each of the following can be considered a simple modification of the original target text. Modifications made include:

1. Changing the verb “loves” to “likes”. This is a relatively minor semantic change but a significant lexical change.
2. Changing the nouns Tom, Mary, Joe to Bob, a cat and Rex. Greatly influence the lexical similarity, but the particular proper nouns used should have little impact on the deep meaning and structure of these vignettes.
3. We also introduce a referent “it”, which makes no change to the meaning of the story but does change the literal similarity.
46 participants were recruited from final year undergraduate cohort studying computer science at Maynooth University, aged between from 20 to 29 (M=21.9). An even number of ratings were obtained for each of the texts being compared. Participants rated the level of similarity they perceived between the target text and the available sources, where the order of presentation of these sources was randomised to control for any potential order effects. The violin plot (below) summarises the responses, where a rating of 5 indicates “highly similar” and 1 indicates “very unrelated”.

**Figure 18: Distribution of human ratings for the four categories of Similarity**

We argue that Dr Inventor’s similarity score is a close approximation of the degree of similarity between these vignettes. The broken lines in the diagram below illustrate the LSA (Latent Semantic Analysis) scores for these texts, which seem to be poor approximations of the human similarity judgements (black solid line). We believe that Dr Inventor’s similarity score (anaSim score discussed later) is a better estimate of similarity than LSA.

<table>
<thead>
<tr>
<th></th>
<th>Literal High</th>
<th>Figurative High</th>
<th>Literal Low</th>
<th>Figurative Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSA doc to doc</td>
<td>0.94</td>
<td>-0.03</td>
<td>0.97</td>
<td>-0.04</td>
</tr>
<tr>
<td>LSA term-to-Term</td>
<td>0.95</td>
<td>0.17</td>
<td>0.97</td>
<td>0.16</td>
</tr>
<tr>
<td>Average Human Rating</td>
<td>71%</td>
<td>62%</td>
<td>60%</td>
<td>54%</td>
</tr>
<tr>
<td>Analogical Similarity (anaSim)</td>
<td>0.477</td>
<td>0.647</td>
<td>0.323</td>
<td>0.291</td>
</tr>
</tbody>
</table>
Analogical Similarity (AnaSim) is a better approximation of human similarity judgements than two alternative versions of LSA.

1. Hidden Similarity and Plagiarism Detection

We briefly illustrate the sensitivity of Dr Inventor’s similarity metrics through their ability to also detect highly similar documents. A dataset was formed from the psychology literature on analogical reasoning, consisting of 36 English texts used on human subjects in various analogy investigations. These texts were translated to different foreign languages and then translated back to English. Due to inaccuracies in the machine translation process, this process creating “plagiarised” versions of each original text (as outlined in Hurley et al, 2016). As can be seen in Figure 20 (a) the translations involving the Arabic and Amharic languages created the least similar documents. We note that these very light levels of similarity between documents (even for the poor quality Amharic round-trip translation, typically result in similarity levels above 0.7). In comparison, the level of similarity underlying the best analogy for any given target is typically in the region 0.3 – 0.45.

Depicted in Figure 20 (b) below are Dr Inventor’s similarity scores for the different categories of plagiarism in the “Sheffield” (Clough and Stevenson, 2011) plagiarism corpus. This corpus consisted of 100 documents, composed of 95 answers produced by 19 different participants to a question & answer task along with the 5 original Wikipedia articles that served as answer to these questions. The corpus contained documents exhibiting 4 different levels of plagiarism:

1. Near copy – text selected from Wikipedia
2. Light revision – as above, with word substitution
3. Heavy revision – as above, with major revisions and re-structuring
4. Non-plagiarism – non-Wikipedia sources

The similarity scores produced for each of the different levels of plagiarism are detailed in Figure 20 (b). As expected the similarity scores reduce for the lighter levels of plagiarism to the Wikipedia-based originals. We note that almost all of the similarity scores produced by this plagiarism corpus are above 0.5 and are thus generally higher than the level of similarity associated with most of the best analogies discovered by Dr Inventor.
Thus, we believe that similarity estimates from Dr Inventor may be useful in detecting “near miss” (almost) literal similarity between documents – in addition to analogical similarity. For the SIGGRAPH corpus containing novel research documents however, we did not see the need to remove any highly-similar documents from consideration.

While Dr Inventor could be used as part of a plagiarism detection system, its current limitation to subjects and verbs would limit its application in its current form. However, it does show potential for identifying at least one form of plagiarism.

5.3 Metrics for Assessing Creative Comparisons

The part of the report describes how Dr Inventor served to identify factors that impact on the comparisons that users find to be most creative. As we will see, this section blurs the distinction between “Analogies” and “Conceptual Blends” per se. We highlight that while “Conceptual Blends” is an umbrella term that includes not only analogies but also “literal similarity” and “mere appearance similarity” (Gentner, 1983). However, throughout our discussion any reference to blends is one who counterpart projection has been identified according to structure mapping theory (Gentner, 1983).

Traditional understanding of analogy-based comparison places the emphasis on the inter-domain mapping and so, better analogies are understood to have larger and more compelling mappings between the two analogs. This is typically built upon with a strong degree of similarity between the relational (verb based) structure and at least some differences between the paired concepts (nouns). As most work in the area has focused on the ability of people to process text, there are no standard metrics for computationally generated analogies and the metrics that do exist are based on predicate calculus representations rather than natural language. A few naturalistic representations have been developed for visual analogies (e.g. Geometric Analogies and Ravens Progressive Matrices, but these are not relevant to our lexically based analogies).

5.4 Analogy Formation and Selection

10 targets were selected using stratified random sampling across the years of the SIGGRAPH corpus. The best source was identified for each, by exhaustively searching each available analog from the corpus. The 10 best analogies were identified, using the AnaSim metric (discussed below). This incorporates semantic factors (relational and conceptual similarity) as well as the Jaccard coefficient to identify analogies that result in a strong mapping (Yalemisew A., et al., 2016).

$$\text{AnaSim} = ((\text{RelSim} + \text{ConSim})/2) \times J\text{Coef}$$
5.5 User Ratings for the Qualities of Creativity

15 experts in computer graphics were recruited and explored these comparisons. Using the Dr Inventor system, users provided ratings for each inter-paper analogy for the following three qualities.

1) This is a Novel or Unexpected comparison
2) This is Potential Useful and Recognizes Gaps in the research
3) This comparison Challenges the norms in this discipline

The ordinal rating data coupled with multiple raters required the use of Krippendorff (Klaus, 2011) inter-rater agreement alpha, returning values between 0.0 and 1.0 with 1.0 indicating maximum agreement. Krippendorf’s alpha for each quality was found to be: Novelty= 0.382, Usefulness=0.26 and Challenge the norms=0.39. While this level of agreement may appear low, we argue that these creativity ratings are still valid firstly because of the relatively large number of rating categories (5), reducing the alpha score. Additionally, creativity is often seen as highly personal and dependent upon users’ expertise and experience. Post-evaluation discussions highlighted why experts gave very different ratings for a few comparisons, focusing on expertise on specific topics. We note that Usefulness showed the lowest agreement with differences in expertise causing disagreement.

5.6 Mapping Qualities

We now explore some qualities of analogies and their relationship to users’ assessments of creativity for these comparisons.

Size of the Mapping (MapSize): We first explored for the expected relationship between the size of the mapping and the ratings awarded by users to each comparison. But surprisingly, a Spearman rank order correlation $r_s = 0.212$ (p=0.279) between mapping size and average rating didn’t show any effect.

Ratio of Mapped Information (MapRatio): To counteract the absolute size of the target, we explored MapRatio which produces higher values as more of the target problem participates in the mapping.

$$MapRatio = \frac{MapSize}{target-size}$$

But again a Spearman rank order correlation was found to be not significant $r_s = -0.006$ (p=0.492).

Jaccard Coefficient (JCoef): We used a Jaccard coefficient to measure the mapping in relation to both source and target analogs (Jaccard, 1901). But again a Spearman rank order correlation did not show any effect $r_s = 0.078$ (p=0.41). So focusing purely on topological qualities of the mapping does not appear to produce the expected effect.

1. Generic Space Qualities

We examine qualities related to the semantic similarity between mapped items, derived from the inter-paper mapping. As these relate not just to one specific space/paper, we associate these with the Generic Space (connecting mapped counterparts).

Conceptual Similarity (ConSim): Conceptual Similarity measures the similarity of paired concepts (nouns) using the Lin metric (Lin, 1998). A Spearman rank-order correlation coefficient revealed a moderate negative relationship between user ratings and the estimate of conceptual similarity $(ConSim)$ $r_s = -0.442$, p=0.09.

Relational Similarity (RelSim): Relational Similarity measures the similarity of paired relations (verbs) again using the Lin metric. A Spearman rank-order correlation showed a weak but positive correlation between user ratings and estimates of relational similarity $(relSim)$ $r_s = 0.430$, p=0.10.

Latent Semantic Analysis (LSA): As a control we also explored for a possible relationship between LSA and creativity ratings of these analogies. Ramscar & Yarlett (2003) previously used LSA in an analogy
model show it to have little application during the critical mapping phase. A Spearman Rank-order correlation between the average creativity rating and LSA score showed no relationship $r_s = -0.6201$ (p<0.05). Thus, while LSA identified no relationship with creative analogies, independent assessment of relational and conceptual similarity appears to provide a more useful avenue for progress.

2. Counterpart and Generic Space Metrics

**Analogue Similarity (AnaSim):** This evaluates the mapping in terms of structural and semantic factors, combining Jaccard’s coefficient with conceptual and relational similarity:

$$\text{AnaSim} = ((\text{RelSim} + \text{ConSim}) / 2) \times \text{JCoef}$$

A Spearman Rank-order correlation between the creativity rating and AnaSim showed $r_s=0.349$ (p=0.16). This is merely suggestive of a mild relationship between AnaSim and creativity ratings.

**Overall Similarity Indicator (OverallSim):** Finally we look at a theoretically driven combination of these metrics. Rather than using the number of metrics we employ an exponential squashing function scaling the number of to the range [0..1] to select analogies with a moderate number of inferences, while comparison offering huge number of inferences will gain little advantage. A Spearman Rank-order correlation between the creativity rating and overallSim showed $r_s=0.1758$ (p=0.315), showing that overallSim was not a very accurate predictor of creative analogies.

3. Qualities of the Blended Space and its Inferences

The next metrics we explore related to the output blended space, incorporating the inferences that are mandated by each comparison.

**Number of Inferences (NumInfs):** Finally, we examined the impact of the Blended space and the number of inferences creativity ratings. A Spearman rank order correlation $r_s=0.286$ (p=0.21) did not show any reliable influence. However, a Wilcoxon paired Sum Rank test revealed that the null hypothesis could not be rejected ($V = 7$, p<0.05). Additionally, a Pearson Product Moment correlation of 0.613 was identified. So, number of inference appears to bear some relationship to creativity analogies.

**Novelty of Inferences (ObservedNovelty):** The raw count of the inferences doesn't address the proper-ties of novelty and quality that are central to creativity (Boden M. A., 2004). In this discussion the quality of inferences is not assessed, due to a lack of domain specific knowledge, amongst other reasons. We assess novelty using an n-gram approach (Yalemisew A. , Diarmuid, Dmitry, & Donny, 2016) derived from the SIGGRAPH corpus.

Firstly, a tri-gram model was constructed from the entire SIGGRAPH corpus of 721,301 triples, with 604,873 distinct triples (ignoring duplicates). As most inferences were found to be novel the corresponding bi-gram model was required, enabling “piecemeal” evaluation of the (relative) novelty of unfamiliar inferences (triples). Thus Subj-Verb, Verb-Object and Subject-Object combinations were evaluated in a piecemeal fashion. For a given analogy producing m individual inferences, the novelty score is calculated as the average novelty scores of the individual inferences.

The Spearman rank order correlation between observed novelty ratings $r_s=0.42$ (p=0.11). While this result was not reliable it does suggest that novelty of inferences is a factor in creativity ratings. Thus we argue that this shows that the novelty of inferences (and not just their quantity) might be a factor influencing users’ perceptions of creativity.

**Novelty Relative to Other Inference (PredictedNovelty):** We also compared inferences against all other inferences generated from all possible analogies from our corpus. We might think of these as inferences likely to occur to someone thinking analogically about the corpus. Dr Inventor explored over 1.3 million analogies producing 225,230 inferences, of which 151,200 were unique (ignoring
duplicates). Tri-gram and bi-gram models were used to estimate the novelty of inferences in relation to all other inferences. However, a Spearman rank order correlation between the novelty of the SIGGRAPH inferences and the user ratings was \( rs=0.048 \) (\( p=0.446 \)) showed that this was not an influencing users’ perceptions of creativity.

4. Combined Attributes of Creative Analogies

We explored linear combinations of the above qualities using principal component analysis (PCA) (Jolliffe, 2002), which is often used to explain the variance within data. A few of the metrics above (eg AnaSim) are already linear combinations of more primitive metrics, this evaluation focuses exclusively on the primitive metrics.

A PCA analysis was conducted and results show that the first principal component accounting for 63% of the variance was formed the following combination of factors:

\[
-0.533 \text{ConSim} + 0.529 \text{RelSim} + 0.485 \text{numInf} + 0.311 \text{PredictedNovelty} \\
+ 0.289 \text{ObservedNovelty} + 0.144 \text{JCoef}
\]

The first and largest principal component indicates (by dint of negative -0.533) that more creative comparisons involve conceptual (noun-based) dis-similarities, as discussed earlier. Conversely, relational (verb-based) similarity appears to be a factor in creative comparisons, as are greater numbers of inferences. The two novelty scores are also important factors in this principal component, helping remove analogies suggesting uncreative inferences. Four principal components account for all variance in this collection.

We believe that the results of this evaluation will allow better identification of creative comparisons in future versions of the Dr Inventor creativity enhancement tool. As previously mentioned, difficulty in attracting user evaluations as well as some difficulties maintaining stable system meant we were not able to roll-out this formula in a subsequent evaluation.

5.7 Evaluating Qualities of Scientific Creativity

During an internal evaluation task that assessed 10 analogies, undertaken during February and March 2016 it was discovered that the quality of analogies presented to Dr Inventor users was not driven directly by the quality or size of the underlying mapping – but rather appeared to be dependent on the inferences that were mandated.

Information was gathered from experts in the domain of computer graphics after using the system. Feedback was sought relating to specific qualities of creativity, arising from the presented comparisons: “What did you think of that analogy?”

1. This is a novel, unexpected or surprising comparison.
2. This is a potentially useful comparison.
3. This comparison challenges the normal understanding of the Target paper.
4. This comparison could help generate new results.
5.8 Feedback on SIGGRAPH Analogies

An evaluation was undertaken on a number of creative analogies from the SIGGRAPH corpus. Potential evaluators who were generally SIGGRAPH authors were contacted directly via email, including specific details on an analogy relevant to them – typically concerning a paper they had written. In response to this direct contact, some users provide their feedback in the form of a reply to the email – rather than using the online system.

We review this feedback as it is a clear indicator of the quality of some of the created analogies – no feedback however was recorded for the majority of the comparisons. We analysed this feedback in terms of the number of words an evaluator used in their reply – this being taken as an indirect indicator of the seriousness with which the evaluator examined the analogy. Additionally, a sentiment analysis of the written feedback was also undertaken using a number of sentiment analysis tools. 204 analogies were identified by internal experts and used to seek the authors’ feedback – though not all were successful in attracting author feedback.

Looking next at the respondents we see that some of these evaluators seemed to readily identify inferences are being particularly useful, while others appeared very reluctant to assert...
that any particular inferences were notably useful. One respondent did not identify any useful inferences from 16 different analogies, while another respondent identified 2 useful inferences from two analogies. On average, respondents identified notably useful inferences 35% of the time.

This evaluation exercise collected 11,334 words of internally based feedback and comments on these 204 analogies, ranging from 8 to 160 words (M=55.58, SD=29.7).

1. Notably Good Inferences

17 evaluators were involved in identifying good analogies, these experts either being project members (primarily from UB and BU) or close associates of project members such as institutional colleagues or professional contact from other organisation. All of these inferences were explicitly identified and recorded separately as being noteworthy. Of the 204 analogies that were identified, 74 (36.2%) seem to involve particularly interesting inferences. Furthermore, 40 of these analogies (54%) involved more than one inference.

These inferences are detailed in other deliverables, but we include a brief sample in Figure 23 below (where blue and orange colours indicate the information originates in either the target or source paper). Analysis of the individual inferences did not yield any obvious patterns.

| The **montage** will **be** the **humans** |
| The **we_our** will **analyze the data** |
| The **surmise** will **sketching the face** |
| The **efficacy** will **allow the user** |
| The **difference** will **editing the albedo** |
| The **result** will **require the pair** |
| The **difference** will **editing the area** |
| The **reflectometry** will **use the corrective that** |
| The **we_our** will **show_predict the motion** |

*Figure 22 Numbers of inferences identified as particularly interesting*

*Figure 23 Sample of the internally identified useful inferences*
The fact that project members (or their close associates) took the time to identify these inferences and record them separately is taken as an indicator of quality.

2. Authors Feedback and Sentiment Analysis

Currently, 24 items of feedback have been received with an average of 31.7 words each, ranging between 8 and 109 words of feedback.

![Figure 24 Number of words in authors repose](image)

The following table summarises the sentiments analysis provided by three different sentiment analysis systems: Sematria Sentiment, Microsoft Sentiment and IBM’s Alchemy Sentiment.

<table>
<thead>
<tr>
<th></th>
<th>Sematria Sentiment</th>
<th>Microsoft Sentiment</th>
<th>Alchemy Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>12</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td>Negative</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Neutral</td>
<td>9</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>% positive</td>
<td>50%</td>
<td>75%</td>
<td>50%</td>
</tr>
</tbody>
</table>
So we see that when feedback was obtained, this generally expressed a positive sentiment. While from this analysis it’s not possible to directly attribute this sentiment to the analogy, we believe this sentiment is a reflection of their views of the system and its creative comparisons.

Table 3 Summary of some indicative contents of Authors feedback on SIGGRAPH analogies

music-driven motion synthesis" is interesting.
not so familiar with material
not sure if this method can be directly applied to hair.
idea looks interesting but little bit difficult to follow.
goal ... unclear to me.
idea 1 makes sense, and actually there was a related work on it "Inverse dynamic hair modeling with frictional contact".
idea make some sense.
could be useful for collision detection between rigid bodies … interesting idea and may be usefully for my future research.
method in that paper would not work with meshes that contain holes, as proposed by the analogy
I’m sure it’s possible to build the hierarchies in a more signal-adaptive manner.
idea is nice
Yes, with part-based templates as prior, we can definitely improve 3D hair modeling. That’s similar to our latest work of AutoHair: Fully Automatic Hair Modeling from A Single Image.
Don’t understand
Don’t understand
basic idea might be relevant and inspiring.
off focus of paper
not likely to work
Yes .. can be improved by many more advanced solutions
no
5.9 Online Feedback on SIGGRAPH Analogies

Once users are presented with a pair of analogies, they were also provided options to rate the inferences. The users were provided binary options to indicate that they like or dislike the inferences. In an effort to collect, further data, users were provided to give written feedback related the analogy they explore.
There are 960 individual pieces of feedback collected from discipline experts including some of the authors of SIGGRAPH papers. This is composed to written textual feedback on a potentially creative analogy, as well as positive (thumbs up 👍) and negative (thumbs down 👎) indications on individual inferences.

Table 4. Statistics: Author feedback on Dr Inventor analogies

<table>
<thead>
<tr>
<th>Written comments</th>
<th>Thumbs up</th>
<th>Thumbs down</th>
</tr>
</thead>
<tbody>
<tr>
<td>218</td>
<td>249</td>
<td>493</td>
</tr>
</tbody>
</table>

1. **Free Text Feedback on Analogies**

Users were given the opportunity to provide verbal feedback on analogies, in response to the following question:

*Did any other ideas occur to you when exploring this analogy? Do you have any other comments or observations?*

Of the 290 evaluation responses received, 248 of 290 (85.55%) attracted some free-text responses. We take this as at least implicit evidence that most of these comparisons were worth considering and were worthy enough of providing some voluntary feedback.

![Figure 27](image)

*Figure 27 Number of words in verbal feedback (b) Visualisation helped my understanding*

71 responses indicated use of the mapping visualisation and 114 reported that they did not use it. As can be seen in part (b) of Figure 22 above, feedback on the visualisation was not very positive, in terms of having a positive impact on users’ ability to understand the analogies. This may be attributed in part to some difficulties in getting the preferred visualisation to work properly on the live system.

### 5.10 Additional Services

1. **Online Text Mapping Service (currently called “2 Map”)**

To facilitate the wider community involved in research on analogical reasoning and conceptual blending, NUIM have created an online service that identifies analogies between submitted texts. This
allows users to submit two passages of text via a simple interface. The combination of text process and cognitive modelling then identifies the largest mapping between these passages.

This differs from the “Try Your Idea” service in that 2-Map identifies analogies between texts, neither of which originates in SIGGRAPH. Because the format of this input didn’t involve SIGGRAPH publications in PDF format, we provided a different (simpler) interface and also hosted the service on Maynooth hardware to avoid complicating the full Dr Inventor service. Some other services hosted by Maynooth used a similar interface, such as the analogies between ICCC publications.

![Figure 28: The online mapping interface for analogy evaluation](image)

The system lists the detailed mapping identified between the two texts. Limited testing of some of the most widely studied analogies (tumour:fortress) has highlighted that Dr Inventor does not accurately identify intended mappings. However, at least some of this limited performance can be attributed to the inclusion of large amounts of superfluous details in the existing stories. Thus, one of the likely applications of this service could be to foster analogy researchers in presenting more focused analogical comparisons, before presentation to users.

Below we present a more focused presentation of the Karla:Zerdia analogy that was developed with the aid of this online mapping service.

**Original Karla the Hawk Story:** Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot a deer instead.

**Original Zerdia True Analogy:** Once there was a small country called Zardia that learned to make the world’s smartest computer. One day Zardia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zardian government realized that Gagrach wanted Zardian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zardia again.

The analogical similarity score between these original texts was 0.198. But using this service we were quickly able to use the tool iteratively to create a more focused version of these stories, forming a more direct and compelling analogy.
Modified Karla the Hawk Story: Karla was a hawk and she had beautifully colorful feathers. It lived in a tall tree. A hunter had some arrows, but the arrows missed Karla. Karla knew the hunter wanted her feathers, so she offered to give some feathers to him. The hunter was so grateful that he pledged never to shoot at Karla in the future.

Modified Zerdia True Analogy: Zerdia was a small country. It made an intelligent computer. Zerdia was attacked by Gagrach. Gagrach owned missiles, but the missiles missed the target. Zerdia realized that Gagrach wanted their computers. Zerdia offered to sell some of its computers to Gagrach. Gagrach was very pleased and it promised never to attack Zerdia again.

Figure 29: Mapping results and Metrics from 2 Map

This deliberately edited analogy yielded an overall similarity Score of: 0.3354. While this similarity score may appear low, we remind the reader that scores around 0.7 are more strongly associated with plagiarism that with figurative analogical comparisons. This score appears to be very consistent with the best analogies discovered by Dr Inventor from the SIGRAPH corpus.

While we intend publicising this 2 Map service to the wider cognitive science community, we are currently improving the reliability of the service and the server and improving the back-end security of this system. Email lists related to cognitive linguistics will be used to publicise this service.

2. International Conference on Computational Creativity (ICCC) Analogies

In an effort to reach a wider audience, especially one that should be more receptive to computational creativity, we applied the Dr Inventor system to a corpus of papers from the International Conference on Computational Creativity (ICCC) 2010 to 2016. Dr Inventor exhaustively explored analogies between papers (from within this corpus), with the best sources-comparisons being retained for each target.
The best five analogs to each target paper were identified for presentation to users. Using an email list associated with the Association for Computational Creativity we attempted to reach a reasonably wide audience of practitioners on Computational Creativity.

Unfortunately, at the time of writing this system had not collected any user evaluations from this endeavour.
6 Evaluation of Creative Analogies

This chapter focuses on reasons that Dr Inventor can be considered a creativity enhancement tool. One of the big achievements that is nearing completion concerns the development of a research project from an idea initially created by the Dr Inventor tool. We then look at two independent evaluation frameworks for creativity, using these to evaluate the creativity of the Dr Inventor system. The chapter concludes with a brief overview of some attempts at Rational Reconstruction of some well-known analogies in computer science.

6.1 The “Flower Paper”

Initiated by Dr Inventor, NCCA developed a new ODE (Ordinary Differential Equation) based method for reconstruction one of hidden structures, which arise from one part of an object (e.g. a flower) occluding other parts of that object. A computational model has implemented the ODE based method in C++ to represent and generate lily flower shapes. Of course, far greater details on this analogy will be found in Deliverable D8.6 “Final version of evaluation report” and only a brief overview is included here.

<table>
<thead>
<tr>
<th>A Research Hypothesis and Project Created by Dr Inventor,</th>
</tr>
</thead>
<tbody>
<tr>
<td>At N.C.C.A., Bournemouth University, UK</td>
</tr>
</tbody>
</table>

A research hypothesis created by Dr Inventor’s led to the following research project. ‘Curve-Skeleton Extraction from Incomplete Point Cloud (2009)’ describes an algorithm for curve skeleton extraction from point clouds, where large portions of data are missing during 3D laser scan. Dr Inventor system has identified ‘Fast Bilateral Filtering for the Display of High-Dynamic-Range Images (2002)’ as analogous, presenting a technique to display high-dynamic-range images, which reduces the contrast while preserving details.

The creative analogy sees both papers focus on the reconstruction of hidden structural information. Paper-1 solves a 3D problem of incomplete vertex data containing holes (caused by self-occlusions during 3D laser scanning). Paper-2 solves a 2D problem in images with poor light management, with under-exposed and over-exposed areas, and light behind the main character.

Proposed solution: The papers are from different Computer graphics domains (Modelling and Image processing) and their methods cannot interpolate with each other. Interestingly, “holes” in the problem are mapped with “areas” in the source paper. Analogous examples from existing work has lead us to the following developments.

Inspired analogy: We have explored new ideas through Dr Inventor to learn; How to rebuild and animate 3D models automatically and reconstruct hidden structure more efficiently. A new idea has been generated after a case-study was carried out from the literature of 18 papers by Dr Inventor. We have found a new method to represent natural flower shape, flower blooming and the decay process, using an Ordinary Differential Equation (ODE)-based surface modelling & simulation technique. Interestingly, the analogy paired “area” with “hole”, normally seen as opposites.

Shape representation of flower is challenging and interesting topic which has attracted the many researchers. The shape of flower consists of a multi-layer architecture (petals, stigma, and stems). Each part of a flower involves a complex geometrical deformation such as bend, stretch, shrink and curl. Various techniques (Data-driven, Sketch-based, Point-based and Image-based) are popular, but face challenges such as the geometry of high fidelity and missing-captured data.

Advantages of our new method: In order to address the existing challenges for the shape representation and simulation of flower, we present a single framework which uses ODE-based surface modelling & simulation technique to solve geometry structural information more efficiently. Our method is very useful for 3D modelling and simulation that creates realistic flower shapes with a small data size.
The following papers were also discovered by Dr Inventor, providing additional information to support the use of ODE based methods for the "flower paper".

- Image-based Plant Modeling (2006)
- Image based Tree Modeling (2007)
- Suggestive contours for conveying shape (2003)
- Projective dynamics (2014)
- Efficient geometrically exact continuous collision detection (2012)
- Example-based elastic materials (2011)

The BU partner explored 26 analogies between SIGGRAPH papers and the additional papers they identified. The following is a sample of these 26 analogies.

A data-driven approach to quantifying natural human motion
Versatile rigid-fluid coupling for incompressible SPH
Brook for GPUs
Performance relighting and reflectance transformation with time-multiplexed illumination

4D Reconstruction of Blooming Flowers
Advanced ODE equation based head modelling
Analyzing Growing Plants from 4D Point Cloud Data
A Flower Growth Simulation based on Deformation

In addition to the initial analogy identified between the two abstracts, we developed the analogy between the two full-papers. This gave much more detail on the comparison including 254 pairings between items from the two papers and accompanied by 270 distinct inferences. Feedback from the BU partner indicated that some of this feedback was very useful leading to the addition of the “Full Paper Analogy” button on the results page – this being available only after a user has explored the initial analogical comparison between two Abstracts.

6.2 Independent Creativity Assessment Metrics (CAQ and CSI)

The study of computational creativity has identified different perspectives on measuring creativity. In this section we examine two of these, using them to illustrate the level of creativity that has been attained by Dr Inventor.

1. Creative Achievement Questionnaire (CAQ)

The Creative Achievement Questionnaire (Carson et al, 2005) is a self-reporting questionnaire for assessing levels of achievement across 10 domains of creativity (Visual art, Music, Creative writing, Dance, Drama, Architecture, Humour, Scientific discovery, Invention and Culinary). The focus of Dr Inventor is on Scientific Discovery. We briefly consider one creative analogy by Dr Inventor that is currently being developed into a research publication.

<table>
<thead>
<tr>
<th>H. Scientific Discovery</th>
</tr>
</thead>
<tbody>
<tr>
<td>__0. I do not have training or recognized ability in this field</td>
</tr>
<tr>
<td>__1. I often think about ways that scientific problems could be solved.</td>
</tr>
<tr>
<td>__2. I have won a prize at a science fair or other local competition.</td>
</tr>
<tr>
<td>__3. I have received a scholarship based on my work in science or medicine.</td>
</tr>
<tr>
<td><strong>4. I have been author or co-author of a study published in a scientific journal.</strong></td>
</tr>
<tr>
<td>__5. I have won a national prize in the field of science or medicine.</td>
</tr>
<tr>
<td>__6. I have received a grant to pursue my work in science or medicine.</td>
</tr>
<tr>
<td>__7. My work has been cited by other scientists in national publications.</td>
</tr>
</tbody>
</table>
Question H.4 of the CAQ asks if the respondent has been “author or co-author of a study published in a scientific journal”. While Dr Inventor was not designed to directly author a scientific paper, it has made significant progress towards this objective. Dr Inventor created an original research idea during an evaluation activity during early 2016. This research idea was subsequently developed into a research project at the NCCA, Bournemouth University, UK. It is hoped (or even expected) that this research project will result in the publication of a conference paper on computer graphics (though probably not a Journal paper). Within the project, this is referred to as the “flower paper” as it concerns a new approach to simulating the depiction of flowers, including their ability to grow and even decay.

We note also that category of the CAQ “G. Inventions” also bares some relevance to Dr Inventor. This contained question (3):

<table>
<thead>
<tr>
<th>G. Inventions</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. I do not have recognized talent in this area.</td>
<td></td>
</tr>
<tr>
<td>1. I regularly find novel uses for household objects.</td>
<td></td>
</tr>
<tr>
<td>2. I have sketched out an invention and worked on its design flaws.</td>
<td></td>
</tr>
<tr>
<td>3. I have created original software for a computer.</td>
<td></td>
</tr>
<tr>
<td>4. I have built a prototype of one of my designed inventions.</td>
<td></td>
</tr>
<tr>
<td>5. I have sold one of my inventions to people I know.</td>
<td></td>
</tr>
<tr>
<td>6. I have received a patent for one of my inventions.</td>
<td></td>
</tr>
<tr>
<td>* 7. I have sold one of my inventions to a manufacturing</td>
<td></td>
</tr>
</tbody>
</table>

While Dr Inventor did not directly create any software, it did contribute the original research idea that led to the creation of a software model by National Centre for Computer Animation - NCCA, Bournemouth University, UK.

2. Creativity Support Index (CSI)

The Creativity Support Index (CSI) is a survey based metric for evaluating the ability of a creativity support tool (CST) to support a user’s creative process. It involves a self-reporting questionnaire conducted in two parts. Firstly, a series of ratings questions assessing the level of importance of different aspects of creativity. This involves answering 12 rating scale questions to assess 6 different factors: Collaboration, Enjoyment, Exploration, Expressiveness, Immersion, and finally Results Worth Effort. Secondly, there is a parried comparison section in which individual factors are compared against each other, involving 15 individual comparisons.
The questions listed above were used to gather feedback data from frequent users of the system, specifically users who had each used Dr Inventor multiple times and given multiple feedback about several different analogies. This identified 13 dedicated users, both from within the consortium and several external users (several of whom were friends or colleagues of consortium members). While calculation of the final CSI results it a little complicated, we summarise our findings here.

CSI uses paired-factor comparison between each of the above six factors paired with every other factor. This comparison is used to estimate the weight assigned to each of the factors by the user (evaluator) of the system resulting a paired-factor count between 0 and 5. A factor with 0 count shows that the factor is not important when using the tool and a value of 5 indicates that the factor is...
crucial. For example the factor “exploration” is very crucial for the respondents as its average count is 4 (Cherry & Latulipe, 2014).

<table>
<thead>
<tr>
<th>Scale</th>
<th>Avg. Factor Counts (SD)</th>
<th>Avg. Factor Score (SD)</th>
<th>Avg. Weighted Factor Score (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Result Worth Effort</td>
<td>3.9(0.94)</td>
<td>15.4 (2.37)</td>
<td>60.6 (20.09)</td>
</tr>
<tr>
<td>Exploration</td>
<td>4.0(0.77)</td>
<td>15.7(1.95)</td>
<td>62.9(14.19)</td>
</tr>
<tr>
<td>Collaboration</td>
<td>1.1 (0.83)</td>
<td>0.0(0.0)</td>
<td>0.0 (0.0)</td>
</tr>
<tr>
<td>Immersion</td>
<td>2.3 (0.9)</td>
<td>14.3(2.19)</td>
<td>33.4(14.38)</td>
</tr>
<tr>
<td>Expressiveness</td>
<td>1.9(1.3)</td>
<td>15.2(1.88)</td>
<td>30.3(21.88)</td>
</tr>
<tr>
<td>Enjoyment</td>
<td>1.6(0.8)</td>
<td>15.4(2.69)</td>
<td>23.6(10.52)</td>
</tr>
</tbody>
</table>

Not surprisingly, the collaboration rating from users was quite low (1.1 out of 5) as Dr Inventor doesn’t support explicit collaboration. Two factors of great relevance to Dr Inventor was “Results worth Effort” (M=3.9) and “Exploration” (M=4.0). Additionally, “immersion” in the task (M=2.3) showed that users were deeply focused on their task of exploring the research literature – and that the tool performed to such a level as to allow immersion in that task.

“Factor score” is the sum of the ratings of the corresponding two questions for each factor (fig 1.4). An excellent score is when the user rate that factor 10 (on a scale of 0 to 10) resulting a factor score of 20. Since Dr Inventor doesn’t support collaboration, the score of collaboration is 0. The factor that gets the highest factor score is again Exploration. The “weighted Factor Score estimate is obtained by the factor Score multiplied by the factor count.

The overall result of the CSI indicated a score of 70.266 (out of 100; SD=11.83) showing that Dr Inventor was successful in supporting the types of creative activity of greatest importance and relevance to its users. Note: CSI scores for individual users ranged from 58 to 93.3 – indicating that some users found that Dr Inventor provided an exceptional level of support for their creativity.

A comparative study conducted by (Cherry & Latulipe, 2014) shows that the average CSI score for Adobe Photoshop being evaluated by novice users is 84.20 (SD=18.84) and AutoDesk SketchBook Express and Bimanual Color Exploration Plugin (BiCEP) get scores of 64.79 (SD=17.06) and 76.52 (SD=16.25) respectively. Despite the difference between the application areas of these systems (artistic creativity) and Dr Inventor, Dr Inventor’s CSI score of 70.2 (SD=11.83) shows that Dr Inventor is successful at supporting creativity.

6.3 Maturity Framework for Computational Creativity

During the course of this project we proposed a new and abstract frameworks for computational creativity (O’Donoghue et al, 2014), one that integrates the quality of creative artefacts into a coherent framework. Truly creative systems should thrive on diversity. But perhaps it might be argued that some creative systems might operate most creatively when there is a distinct lack of diversity in its inspiring set. That is, does the creativity of a creative system weaken as more variation is given to its inspiring set? Can a system still function creatively when exposed to the full breadth of diverse artefacts that one would expect to find a real world “inspiring set”?

We believe that a general purpose means of estimating the maturity of a computationally creative system needs to assess the quality of its creative artefacts. We offer the perspective that artefacts of
sufficient creativity must possessing sufficient novelty and quality to be deemed creative, with the outputs directly produced possessing these qualities - Direct Computational Creativity in Figure 27 below.

A higher degree of maturity will be required of computationally creative systems before their computationally created artefacts can be successfully added to the “inspiring set” such that this results in an improvement to the overall creativity of the system. We do not believe that Dr Inventor (or most other systems) has reached that level of maturity. There is a deceptively large distance between the DCC and DSC requiring at least two distinct

1) Ensuring computationally created artefacts poses sufficient quality and novelty

2) Ensuring computationally creative system are written in such a way as to take advantage of a continually changing “inspiring set”, potentially requiring modification to the mechanisms that assess creative qualities like novelty and quality/usefulness.

The big challenge for Dr Inventor in progressing from DCC to DSC concerns the quality (or usefulness) of its creative artefacts. That is, we argue that Dr Inventor performs an in-depth assessment of the novelty of its creative artefacts – focusing as it does on the creative inferences. However, it does not possess an ability to assess the usefulness (or even validity) of its creative artefacts.

The community of researchers on computational creativity are actively engaged in research that aims to create computationally creative systems, whose outputs are creative artefacts (music, images, hypotheses, etc). Can we envisage a future where computational systems attain the same level of creativity as our current researchers on CC? Our maturity framework identifies two additional levels of maturity for computationally creative processes. Indirect Computational Creativity (ICC) where the created artefacts are themselves produces of creative artefacts – followed by the self-sustaining extension of that level.

Dr Inventor can be seen as operating at the DCC level – as can most current systems on computational creativity. While most computationally creative system can produce creative outputs, we are not aware of any system whose outputs are of sufficient quality that adding such an output to its “inspiring set” would actually improve its creativity. The biggest hurdle for Dr Inventor towards achieving this
level of creativity concerns the quality of its Inferences (and its output space). We are not aware of any work on computationally creative systems that aim to output a system that itself has creative abilities.

6.4 Rational Reconstruction

While rational reconstruction will be addressed in other deliverables, we briefly visit this topic. Here, we wish to assess the ability of Dr Inventor to re-enact past instances of creativity using only the resources that would have been available at that time.

1. Simulated Annealing

A highly cited paper that clearly highlights its analogical origins is Simulated Annealing (Kirkpatrick et al., 1983). "A detailed analogy with annealing in solids provides a framework for optimization of the properties of very large and complex systems. This connection to statistical mechanics exposes new information and provides an unfamiliar perspective on traditional optimization problems and methods”.

However, looking through the literature on (metallurgical) annealing and also the literature on network systems (or neural networks) shows that Dr Inventor would (probably) not identify the analogy underlying simulate annealing. This can be somewhat attributed to the way that the temperature of the metal is described, generally being presented as a quality of the entire metal, rather than being associated with (atomic-like) parts of that metal. Related to this is the fact that the atom level of metals may not be described in a way that makes it easy to align metallic atoms with individual parts of a network. Furthermore, the pathways supporting the movement of energy between atoms are not necessarily discussed explicitly, but are essential to developing the correct analogy. Thus, while the documents we explored did not directly support reconstruction of the analogy underlying simulated annealing – we believe that Dr Inventor is theoretically capable of generating this comparison if it was supplied with documents focusing on the correct information.

2. PageRank

The well-known PageRank algorithm underlying the Google search engine can be seen as arising from an analogy between the structure of the web (WWW) and citation analysis. "The citation (link) graph of the web is an important resource that has largely gone unused in existing web search engines" (Page and Brin, 1998). Furthermore, PageRank also borrows some mathematical concepts from the mathematical domain on “the random walk of a graph”.

Attempts to reconstruct some or all of this analogy were also unsuccessful, partly as most of the documents on citation analysis were very detailed. We did not encounter any documents that described (say) a single web-page with links and a single publication with citations in an appropriate manner. While recreating this analogy might be technically possible, we do not believe it is likely that Dr Inventor would re-discover this analogy unless documents focused on specific aspects of the two concepts. Without the core analogy between a web-page and a publication, the mathematics of the random walk contributes nothing.

6.5 Other Good Comparisons

The following gives an example of the kind of detailed qualitative feedback received for one specific analogy created by the Dr Inventor system.
1. Hair Modelling

This analogy centres on hair modelling and qualitative feedback received from user E.C. on 13th Nov. 2016, who informally rated this as a “Highly useful idea”. (Note that the other three analogies were variously rated between “Useful idea” and “Less useful idea” by this user).

**TITLE OF TARGET PAPER:** Single-view hair modelling for portrait manipulation (SIGGRAPH 2012)

**TITLE OF SOURCE PAPER 2:** Texture-lobes for tree modelling (SIGGRAPH 2011).

“The source paper gives another way to solve the problem of the target paper. The patches (lobe-geometry) based method is also suitable for hair reconstruction somehow. Hair modelling from single-viewed image is lack of depth information. But we can still cut the hair region into several meaningful parts like the lobe-geometry in the source paper. The skeletal structure idea can be implemented in the target paper by selecting representative 2d hair strands. And if a hairstyle database is available to training, the 3d hair model reconstruction can be done by incorporating skeletal structure with hair region partition.”

This shows that Dr Inventor discovered a comparison, whereby a method described in one paper may be highly applicable to a (very) different problem area. Hair modelling is normally not seen as being similar to tree modelling.

2. Conclusion

Our analysis point to some mixed findings. Firstly, we note that actually getting users to use the system was a major challenge, despite the number of outreach activities undertaken. Continued efforts to foster use of the system consumed a large amount of time and energy. In the end, one of the main sources of external feedback was a large emailing campaign incorporating detailed information related to each specific user.

The feedback that was received however was generally very positive. The best evidence of this is the results of the CSI evaluation, directly assessing the level of support provided by Dr Inventor for the creative task of its expert users. The parts of the CSI that relate most directly to Dr Inventor (“Result Worth Effort” and “Exploration”) scored very highly leading to the overall level of creativity support at over 70% was very pleasing and compared favourably to some professional tools supporting users creativity at more well established tasks (image creation).

Additionally, the feedback obtained from a relatively small number of users that made significant use of the system was also very positive. Perhaps the most convincing evidence is the duration of the average visit to Dr Inventor at over 9 minutes per user session (see Deliverable D8.4).
7 Conclusion

7.1 Future Work

In this section we briefly look at some specific modifications and improvements that we believe, might improve the quality of the analogies/blends identified by Dr Inventor.

Despite the significant efforts made to derive a “deep” semantic representation of the contents of each document, Dr Inventor is still heavily influenced by the surface level presentation of information in each document. Seemingly small differences in the presentation of a text, especially those impacting the dependency tree, can have a large impact on the generated ROS and thus on the analogies that are identified. Because of its implicit dependency upon the presentation level of a document, we believe that Dr Inventor may best be thought of as a discourse mapping system.

1. Quality of the Analogies

While Dr Inventor made significant achievements towards identifying analogies between lexical documents, we need to make a point that work in this direction is far from complete.

One of the most significant limitations in identifying deep analogies between text-based data, remains the quality of analogies that are typically identified. While Dr Inventor successfully identified many analogies, it is not necessarily the case that these were either the best analogies possible, nor were they necessarily the same ones that would have been identified by people. The quality of the analogies that are identified is still a major area for improvement. For example, if we look at the Tumor:Fortress problem, we see that the analogy that was identified by Dr Inventor differs significantly from any interpretation generated by people.

Tumor Problem:

Suppose you are a doctor whose patient has an inoperable tumor in their stomach. It is impossible to operate, but the tumor must be destroyed or the patient will die. But an X-ray can be used to destroy the tumor. If strong X-rays reach the tumor it will be destroyed. Unfortunately, all the healthy tissue that the X-rays pass through will also be destroyed. At lower intensities the X-rays are harmless to the healthy tissue, but will not affect the tumor. What type of procedure might be used to destroy the tumor with the rays, while avoiding destruction of the healthy tissue?
Fortress-1 Dispersion Attack:
A small country fell under the iron rule of a dictator. The dictator ruled the country from a strong fortress. The fortress was situated in the middle of the country, surrounded by farms and villages. Many roads radiated outward from the fortress like spokes on a wheel. A great general arose who raised a large army at the border and vowed to capture the fortress and free the country of the dictator. The general knew that if his entire army could attack the fortress at once it could be captured. His troops were poised at the head of one of the roads leading to the fortress, ready to attack. However, a spy brought the general a disturbing report. The ruthless dictator had planted mines on each of the roads. The mines were set so that small bodies of men could pass over them safely, since the dictator needed to be able to move troops and workers to and from the fortress. However, any large force would detonate the mines. Not only would this blow up the road and render it impassable, but the dictator would then destroy many villages in retaliation. A full-scale direct attack on the fortress therefore appeared impossible. The general, however, was undaunted. He divided his army up into small groups and dispatched each group to the head of a different road. When all was ready he gave the signal, and each group charged down a different road. All of the small groups passed safely over the mines, and the army then attacked the fortress in full strength. In this way, the general was able to capture the fortress and overthrow the dictator.

The mapping doesn’t one of the central correspondences of the Tumor <-> Fortress analogy. Clearly, this does not correspond exactly to the analogy as it is understood by human “interpreters”. So even though Dr Inventor does perceive an analogy between these stories, it differs in crucial ways from the comparison as understood by people. The mapping suggests that Dr Inventor was overly swayed by the two instances of the verb “be” mapping to itself, with this similarity overshadowing the desired comparison. Furthermore, the metrics (below) show that this interpretation of the comparison did not produce any inferences. It is these (missing) inferences that are the main point of this analogy.

<table>
<thead>
<tr>
<th>Source Word</th>
<th>Target Word</th>
<th>Word Sim Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>be</td>
<td>be</td>
<td>1.0000</td>
</tr>
<tr>
<td>intensity</td>
<td>way</td>
<td>0.5761</td>
</tr>
<tr>
<td>pass</td>
<td>give</td>
<td>1.0000</td>
</tr>
<tr>
<td>[unknown]-0</td>
<td>signal</td>
<td>-1.0000</td>
</tr>
<tr>
<td>tumor</td>
<td>mine</td>
<td>0.0000</td>
</tr>
<tr>
<td>be</td>
<td>be</td>
<td>1.0000</td>
</tr>
<tr>
<td>tissue</td>
<td>body_them</td>
<td>0.4840</td>
</tr>
<tr>
<td>reach</td>
<td>pass</td>
<td>1.0000</td>
</tr>
<tr>
<td>be</td>
<td>have</td>
<td>0.3758</td>
</tr>
</tbody>
</table>

Mapping Size 9
Similarity Score 0.6795
Relational Similarity Score 0.8752
Conceptual Sim Score 0.3533
<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard Coefficient</td>
<td>0.1800</td>
</tr>
<tr>
<td>Overlap Coefficient</td>
<td>0.6923</td>
</tr>
<tr>
<td>Mapping Ratio</td>
<td>0.1957</td>
</tr>
<tr>
<td>Analogical Similarity</td>
<td>0.1223</td>
</tr>
<tr>
<td>InferenceSize</td>
<td>0</td>
</tr>
<tr>
<td>Overall Similarity</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

The style in which the original documents were written still has a large impact on the quality of the identified analogies.

2. **Improvements in Knowledge Extraction**

While we believe Dr Inventor represents a great step forward in terms of semantic processing of language, there are some issues that could improve its results.

*Representing negation*: The current lexical parser driving the ROS generation process performs poorly at explicitly representing negation. This might help improve the quality of ROS and subsequent analogies. Indeed, detecting contradictions between inferences and previously held beliefs might also help improve the creativity of the model.

*Incomplete Triples*: Many triples are still incomplete. Recent advances in the Dr Inventor text processing pipeline helped reduce the number of incomplete triples, but at the cost of each triple greatly expanding in length (see “4.2 Representing Causal Relationships”). Unfortunately attempts to incorporate this casual information into the mapping process were not successful. Additional or alternate approaches such as RelEx (Fundel et al., 2007) dependency parser might help reduce these incomplete triples.

*Active/Passive Voice*: The subject and object roles of verbs are currently reversed for many sentences expressed in the passive voice. Consequently the mapping process maps roles flexibly, using structure and semantics as an influence to guide discovery of the optimal mapping, but results could be expected to improve with better treatment of the active and passive voice.

*Umbrella Terms and Synonyms*: In discursive text writers frequently use selected terminology to about repetition of specific terms. Even in the SIGGRAPH corpus, at some points in a text the authors may talk about the *algorithm*, the *code* and the *program* as interchangeable terms. While in other parts of the same text these terms might have significantly different meanings. It seems an open problem to distinguish between co-references involving dissimilar terms, while at other stages similar terms may refer to distinct concepts. This may be somewhat related to the notion of “roles” and “role mapping” discussed most recently by (Leuzzi and Ferilli, 2017).

This issue might be somewhat mitigated by Word Sense Disambiguation systems like Babelfy (Moro et al., 2014) that spot and link distinct lexicalizations of the same concept (e.g. car, automobile, etc.) to the same concept identifier, thus enabling a better management of the lexical variability within a text. Concepts that are identified by Word Sense Disambiguation and Entity Linking systems like Babelfy are connected each other by a rich set of relationships that could help to measure their semantic similarity or the existence of hyponymy relations.

Other improvements that might be considered include Named Entity Recognition that is explicitly handled by some parsers, as this might assist with some of the author names, algorithm names and other concepts that pervade scientific publications.
3. **Better Word Similarity Estimates**

The semantic component of Dr Inventor occasionally returned zero results for some paired terms. One means of overcoming this and increasing its coverage of comparisons is to use multiple metrics (Lin, Wu&Plamer and others), thereby allowing computation of some semantic comparison value for a wider range of terminology. Perhaps even a non-linear combination of multiple established metrics would help address the prevalence of this zero-similarity problem.

Currently, Dr Inventor's semantic component does not commit to a single interpretation of a term involved in an analogy. This decision was made to allow maximum flexibility among the comparison and partly arose due to the poor coverage of scientific terms in WordNet. In an ideal world, Dr Inventor should be able to identify a single interpretation of each word in a document.

4. **Scientific Discourse Ontology (SDO)**

The biggest "missed opportunity" from this project involves failure to make better use of the SDO categories. While various combinations of both lexical and SDO subsets of papers were generated, we were not in a position to get user feedback to compare these categories due to the slow uptake of user interactions.

5. **Automatic Document Summarisation**

Automatic document summarisation offers one means of creating more compact representations of papers, from which to explore analogies. This may enable some of the more voluminous SDO categories to be summarised and potentially even normalise their impact on the resulting comparisons.

Summarisation might be seen as ever more important for larger documents, given our theoretical investigation (Section 4.4) showing the super-exponential growth in the number of possible topologies for a graph containing N edges. Testing seems to be one approach to determining the best summarisation length that leads to the identification of the best initial analogies between text-based documents and we may additionally expect different sub-disciplines to offer even greater potential for improved operation.

6. **Alternative Knowledge Representation and Data Sources**

**FrameNet:** An alternate approach to identifying analogies between text-based documents would be to adopt the knowledge representation approach of FrameNet. The FrameNet approach explores more semantically funded relationships within lexical information, thereby offering the potential. The original text of an example sentence was:

"It can be hoped that the Spanish Prime Minister Felipe Gonzalez will draw the right conclusions from his narrow election victory Sunday. A strong challenge from the far left, the communist Izquierda Unida, failed to topple him. He should consider his victory a mandate to continue his economic reforms and not a demand that the move further left. If he follows the correct path, he may be able to look back on this election as the high water mark for left opposition".

Leading to the FrameNet annotated representation⁴.

---

⁴ https://framenet.icsi.berkeley.edu/fndrupal/fulltextindex
NELL: (Never-Ending Language Learning): Other projects like NELL (Mitchell et al., 2015) from CMU also read lexical information and represent its meaning in a form loosely akin to our ROS. It has acquired a knowledge base with over 80 million confidence-weighted beliefs (e.g., `servedWith(tea, biscuits)` and text box below.

“Additionally, it has learned to reason over these beliefs to infer new beliefs, and is able to extend its ontology by synthesizing new relational predicates”

Microsoft Concept Graphs: Microsoft released its Microsoft Concept Graphs for short text understanding (Wu et al., 2012; Cheng et al., 2015) in late 2016, which it describes as a “Single Instance Conceptualisation” supporting disambiguation between polysemous terms. They describe it as a universal, probabilistic taxonomy that is “more comprehensive than any existing ones”. This might help improve systems like Dr Inventor in the future, especially as it too is based on Neo4j graphs (like our ROS).

IBM Watson Natural Language Understanding and AlchemyAPI: Dr Inventor previously used the IBM Watson API to assist in evaluating novel inferences. During the course of this project that service moved from Watson to AlchemyAPI. Very recently however, this service has now moved back to IBM as Watson Natural Language Understanding (NLU). Watson NLU may well provide additional services assisting in the evaluation of inferences, especially inferences arising from comparison between two papers from different research disciplines (say computer graphics and chemistry).

Availing of some of these resources may help improve the quality of the semantic information derived from the originating documents.

7.1.6.1 Mapping Improvements

Introducing the VF2 based graph matching algorithm resulted in great improvements to the quality and the speed of the mapping process. Subsequent parallelisation of this algorithm further improved its speed and opened up the possibility of a more scalable creativity engine.

However the quality of analogies is still dependent on WordNet-based similarity metrics, which have numerous limitations. Not least is the fact that none of the terms found in current ROS...
graphs are tagged during lexical analysis by their synset number. Evaluating mappings with greater precision in terms of the semantic contents of inter-ROS mapping would be a great improvement.

Texts, including scientific ones, are often written so as to avoid over-repetition of some central terms. A computer graphics paper may present a new algorithm, variously referring to it as “the algorithm” and also “the program” (there may often not be identified explicitly as co-reference terms). For a complete mapping to be formed, any source paper would (for best results) also employ two terms to refer to an analogous concept. Some analogy papers talk about “roles” with each analogy where multiple lexical terms refer the same semantic concept, akin- to the LISA model (Hummel and Holyoak, 1997). This is a very challenging problem as distinguishing between implicit references and reference to essential distinct concepts is frequently very unclear.

7.1.6.2 Inference Assessment

One of the biggest and arguably an open ended challenge for computational scientific creativity, up concerns evaluating inferences for their quality - as opposed to assessing them for their novelty.

The only relevant work are aware of in scientific analogising is a model for qualitative reasoning within physical domains called Phineas (Falkenhainer 1987, 1988, 1990). This was based on Ken Forbus's "Qualitative Process Theory" (Forbus, 1984). Phineas too was a model of analogy built for the SME analogy mapping model, being particularly focused on the post-mapping process of "Verification". While Phineas used very detailed symbolic representations of physically derived information, this had the impact of making it extremely tied to the domain of qualitative physics. The Phineas approach offers the possibility of modelling the kinds of “mental simulation” that often precede any actual detailed reasoning.

Relate to this might be the use of “reason maintenance systems” and “Truth maintenance systems” to support creation of explicit belief spaces. Thereby, contradictory belief spaces generated by Dr Inventor could be separately maintained, with the possibility of detecting “transformational creativity” that diverges from existing belief systems.

7. Conclusion

Overall, the Dr Inventor system showed that computational means can be fruitfully used to both assist and enhance the creative activities of practising scientists. Surprisingly, one of the biggest challenges appears to be encouraging potential users to adopt the system to and to explore the comparisons that have been identified. While the Dr Inventor team expected much difficulty in collection users data and feedback, the level of un-interest that was apparently displayed was extremely surprising – at least to some of the project members. Some possible reasons for this include: difficulty in visualising the analogy and a web interface involving a significant cognitive work-load on the user. While not all analogies were rated positively, this doesn’t appear to be a factor in the low adoption rates. However, the positive feedback of those that used the system is taken as a positive indicator of the system ability and of this approach to enhancing the professional creativity of practising scientists in an age of exponentially growing academic output.
8 References


## Appendix 1 – Abbreviations and acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CET</td>
<td>Creativity Enhancement Tool</td>
</tr>
<tr>
<td>CST</td>
<td>Creativity Support Tool</td>
</tr>
<tr>
<td>LCS</td>
<td>Largest Common Subgraph</td>
</tr>
</tbody>
</table>