Conceptualising Computational Creativity: Towards automated historiography of the research field

Vid Podpečan[1], **Nada Lavrač**[1], **Geraint Wiggins**[2,3], **Senja Pollak**[1,4]
[1] Department of Knowledge Technology, Jožef Stefan Institute, Ljubljana, Slovenia
[2] Computational Creativity Laboratory, AI Lab, Vrije Universiteit Brussel, Belgium
[3] School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[4] USHER Institute, University of Edinburgh, Edinburgh, UK
{vid.podpecan,nada.lavrac,senja.pollak}@ijs.si; geraint.wiggins@vub.be

Abstract

This paper reports on the progress towards constructing automated historiography of the research field of Computational Creativity (CC). The analysis is based on papers published in the Proceedings of International Conferences on Computational Creativity in eight consecutive years since 2010. This paper extends our earlier work by proposing an approach to CC field analysis that facilitates the automation of CC conceptualisation.

Introduction

Computational Creativity (CC) is concerned with engineering software that exhibits behaviours that would reasonably be deemed creative (Boden, 2004; Colton and Wiggins, 2012). As for every other research community, it is crucial for the CC community to analyse its research topics, applications and the overall progress of the field with the goal of CC field conceptualisation.[1]

Loughran and O'Neill (2017) have studied the CC domain by analysing its conferences and proceedings, where—as they acknowledge—conceptual categorisation was conducted subjectively, through a review of each paper. In contrast, the aim of the research presented in the current paper is to provide an semi-automated analysis of the field as it develops, with the expectation that this may be used in the future for automated construction of the historiography of CC research, which can substitute or complement manual analysis of the research field. Our long term vision is to provide a system, which would be fully automated and available online to the CC community for its analysis and promotion to a wider public.

The conceptualization of the CC research field has been studied already in our past research, where a mixture of text analysis and clustering methods was used (Pollak et al., 2016). In this paper we report on further work in this direction, complementing the previous study by introducing an extended set of methods and by analysing papers published in additional ICCC proceedings. We show how the extended set of methods can be used to support the understanding of the conceptual structure of the field as represented by the papers presented at its annual International Conference on Computational Creativity (ICCC).

The paper is structured as follows. First, we briefly review the previous attempt to address this question. Next, we describe the data used in the study, followed by the section in which we explain the methodology that (a) supports topic analysis through diachronic clustering, (b) uses a contemporary visualisation method, and (c) involves relatively little human intervention, to the extent that can be fully automated in the future. We present the results of this methodology and explain the achieved conceptualisation.

Experimental data

We used the ICCC corpus presented by Pollak et al. (2016) constituting of the articles from the proceedings of the 2010–2015 International Conferences on Computational Creativity, and complemented it with the papers from the years 2016–2017. The text files were converted from PDF to TXT and the bibliography sections were removed. Our corpus consists of 340 articles in total (see Figure 1).[2]

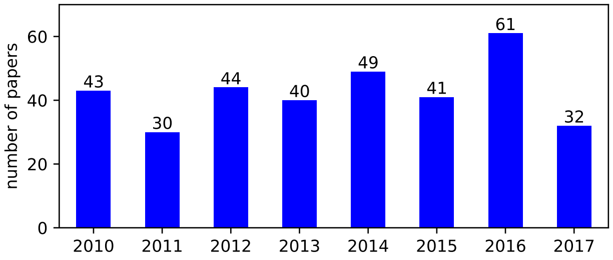

Figure 1: Numbers of papers in the ICCC Proceedings.

[1] We use the term conceptualisation in alignment with its standard use in information science, where conceptualisation is defined as an abstract (simplified) view of some selected part of the world, containing the objects, concepts, and other entities that are presumed of interest for some particular purpose and the relationships between them (Gruber, 1993; Smith, 2003).

[2] Note that there might be minor differences between the number of articles in the corpus and the actual proceedings, since for 2010–2015 the corpus was manually collected and we cannot exclude human mistakes, while for 2016 and 2017 it has been crawled automatically, but we have noticed a few document duplicates.

Previous results

In our previous work we performed domain conceptualisation by applying semi-automated, user-guided clustering using a topic ontology construction tool OntoGen (Fortuna, Mladenič, and Grobelnik, 2006). The resulting corpus-based categorisation of the field of Computational Creativity, presented in detail by Pollak et al. (2016), identified the following main subdomains of Computational Creativity: Musical, Visual, Linguistic creativity, Games and creativity, Conceptual creativity, as well as the domain of Evaluation, which was added after a manual query used in the active learning approach to topic ontology creation. For several subdomains, subcategories were detected at a lower level, including Narratives, Poetry, Recipes and Lexical creativity as subdomains of Linguistic creativity.

Proposed Domain Conceptualisation Methodology

The main ingredients of the extended methodology are: detection of CC topics by document clustering, enrichment of the analysis by cluster visualisation, and performing clustering incrementally on different datasets, starting with the first edition of ICCC 2010, and finally using the entire ICCC proceedings data set (2010–2017) in the final analysis, thus mimicking the continuous automatic analysis support that we aim to make available to the community in the future.

Data cleaning and preprocessing

First, we have performed a number of preprocessing steps in order to make the data suitable for the analysis. One by one, the articles from the ICCC corpus were sent to the following pipeline to obtain lists of tokens:

1. decode all characters from UTF-8 to ASCII using the Unidecode library[3] in order to remove some of the artifacts introduced by the PDF-to-text conversion;
2. split sentences using the Punkt tokeniser (Kiss and Strunk, 2006);
3. expand contractions;
4. word tokenisation using the Treebank tokeniser (Marcus et al., 1994);
5. token filtering to remove tokens of length less than two, unprintable tokens, numbers, and non-alphanumeric tokens;
6. lemmatisation using the LemmaGen lemmatiser (Juršič et al., 2010);
7. adding bi-grams and tri-grams;
8. removing stopwords.

In spite of elaborate automated preprocesing to remove PDF-to-text artefacts, several smaller issues such as hyphenation, ligatures etc. remain and can be observed in some of the visualizations. For example, the character sequence *tion* is a common ending of several hyphenated words and thus appears as an important term in several wordclouds.

Diachronic paper grouping

This research aims to provide a methodology for continuous, automated historiography of the field. In this setting, after each conference, the editors would upload the papers to the system, and the clustering (topic identification) would be automatically produced. The resulting information, essentially a set of topological representations, can then be used computationally to create descriptions of all or part of the field.

For this reason, we group the papers cumulatively by year, starting with the first edition of ICCC, year y1=2010, then adding the next year's proceedings to the corpus in year 2 (y2=2010–2011), and so on. The latest set of documents consists of all the available papers (y8=2010–2017).

Clustering

In order to perform document clustering, vectors of tokens as returned by the preprocessing pipeline described above were first transformed into tf·idf vectors (Term Frequency·Inverse Document Frequency: Salton and Buckley, 1988). This was followed by Latent Semantic Indexing (LSI) (Deerwester, 1988), which performs singular value decomposition and keeps only the largest values thus effectively reducing the dimensionality by several orders of magnitude and reducing noise.

In general, determining the optimal number of target dimensions when performing LSI is still a challenge. For a real world sized corpus with e.g., 10^5 documents, a number such as 300 is considered as appropriate (Bradford, 2008). Taking into account that our corpus consists of only 340 articles, we have set the desired number of dimensions to 10 after a series of experiments where the *silhouette score* (Rousseeuw, 1987) was measured for a different number of target dimensions and a different number of clusters. When the number of LSI dimensions was around 10, the silhouette score did not show anomalous trends such as monotone increasing or decreasing and visualization of the corpus revealed clearly visible groups of data points, which

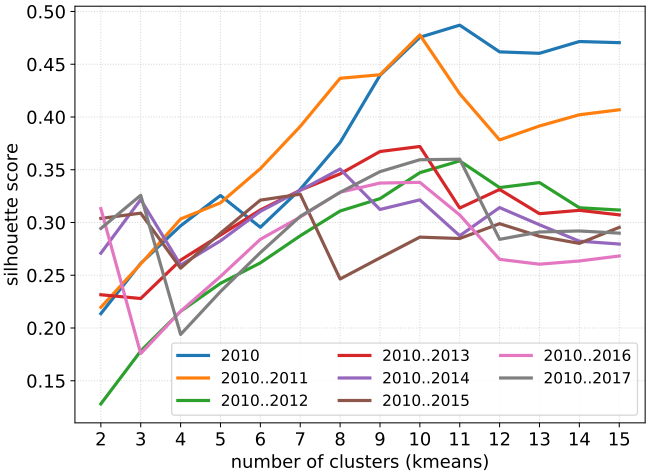

Figure 2: Silhouette scores for different values of *k* and different (cumulative) document sets.

[3] Unidecode is based on hand-tuned character mappings that also contain ASCII approximations for symbols and non-Latin alphabets.

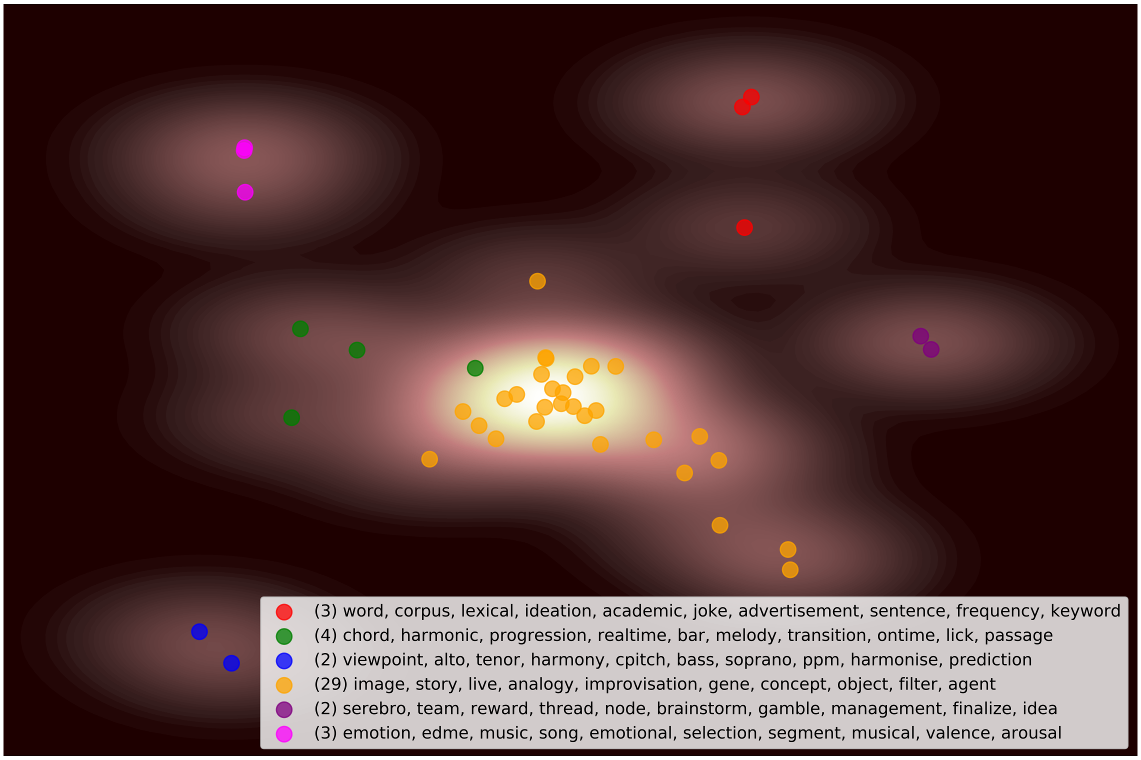

Figure 3: A visualisation of the ICCC 2010 proceedings papers clustered into 6 clusters.

indicates that such setting is appropriate for revealing the structure of the corpus and for reducing the noise introduced by text extraction.

Since we are interested in fully automated methods, we experimented first with the DBSCAN clustering algorithm (Ester et al., 1996) that does not require the number of clusters set as a parameter. However, the results were poor and the algorithm was not able to find dense regions in the data. Therefore, we resorted to the *k-means* clustering algorithm. For manually selecting the *k* parameter (the number of clusters), the user can rely on the visualisation of the document space and expert knowledge of the domain. In addition, we have evaluated the silhouette score for 2 to 15 clusters (see Figure 2) in order to investigate whether the optimal number of clusters can be determined automatically. We compared the results of manual and silhouette-based *k* setting, and decided to focus on the results with manually set *k* as they led to more meaningful interpretations. We defer automatic parameter setting experimentation to future work.

Finally, the results of clustering were also used to automatically extract keywords. For each cluster, the top *t* terms (tokens) of the mean tf·idf vector (centroid) are collected and presented to the user. The terms can be used to identify main topics, the diversity of the cluster, detect outliers and evaluate whether the number of clusters is appropriate.

Visualisation

We have devised a visualisation methodology to support historiographic analysis of a given domain described by a set of documents and a timeline. The methodology consists of (a) combining the results of clustering with a 2D projection of the document space, (b) wordclouds and (c) composition of the result of (a) from different time points into a video clip.

2D visualisation First, the documents are preprocessed and LSI vectors are computed and clustered as described in the preceding subsection. Then, an Isomap (Tenenbaum, Silva, and Langford, 2000) projection is computed which yields 2D coordinates for each document. Using Isomap results we draw a scatterplot where each point represents one document. In addition, the cluster index is used to assign colours to points, while top-weighted centroid terms (keywords) are used for cluster summarisation[4]. On top of that, we use a 2D kernel density estimation to compute the shading of the scatterplot background. This visualisation is shown in Figure 3.

Wordclouds In addition to the 2D map of the corpus we also compute and display wordclouds that can help in identifying the keywords and topics of the selected document set (which can be either the current year or a cumulative set of all years up to the current time point). Figures 4 and 5 show wordclouds for the first and the last year of the ICCC

[4] In all the presented figures only 10 keywords per cluster are shown due to limited figure width.

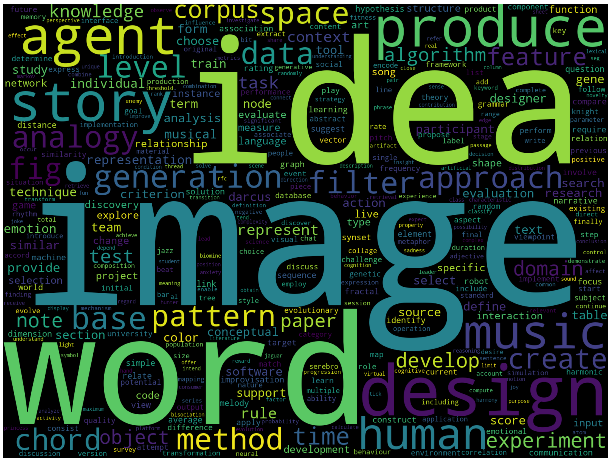

Figure 4: A filtered wordcloud of the 2010 ICCC articles.

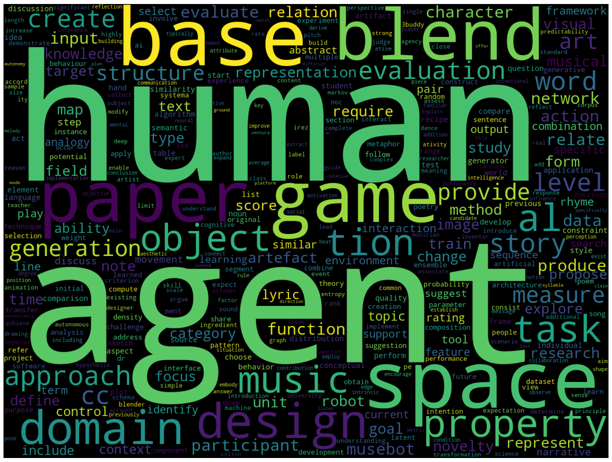

Figure 5: A filtered wordcloud of the 2017 ICCC articles.

proceedings corpus with the following most frequent general terms manually removed prior to wordcloud drawing: *creativity, creative, model, process, computational, result, generate, concept, set, figure*.

Animation The described 2D visualisation procedure can be used to create animations of the changes of the document space through time. Such animation can be used to follow the development of topics through time, observe merging and splitting and detect trends. To generate a movie, the visualisation procedure is applied sequentially to a growing collection of documents. In each time step, new documents are added and a new image is produced. Finally, the pictures are merged in a video clip and the crossfade effect is applied to smooth the transitions from one image to another.

We have also implemented a modification that enables tracking of topics/clusters. By default, colours for clusters (data points) are selected randomly. This is sufficient for single images but may introduce confusion when several images are merged into a video because a cluster of a certain colour is not necessarily related to a cluster of the same colour on the next video frame. Therefore, we have implemented a heuristic approach that works as follows. For each time step we compare the current clustering keywords with the previous ones. Whenever a high level of similarity between two ranked keyword lists is detected we assume that this is the same cluster so the same colour will be used in the current image. In addition, we change the shape of the scatterplot points to allow for visual detection of such clusters. The similarity between two ranked lists is computed using the Rank-biased overlap algorithm (Webber, Moffat, and Zobel, 2010) and was found to reliably detect similarities between ordered lists.

Results: ICCC Topics Across the Years

We analysed the results for different values of k and different sets of years. For example, if we input the papers of the first edition of ICCC in 2010, and make a single split into 2 clusters (Figure 6), we see that the documents are grouped into *Musical creativity*, with the keywords *music, melody, harmonic, song*, while the other cluster comprises all other themes. For deeper understanding, we can see the title of the documents, and the paper with the title “User-Controlling Expressed Emotions in Music with EDME” explains the keywords *edme* and *emotion* in the cluster name. The clustering probably illustrates the familiar problem of disjoint terminology between musical creativity papers and others.

If we set k higher, we can get a more realistic topic overview. For instance in Figure 3 (same document set, split into 6 clusters), *Musical creativity* can be observed across several clusters, one related to the modeling of harmony (blue), while others cover the papers related to emotions and music (pink in upper left corner with documents “Real-Time Emotion-Driven Music Engine”, “Automatic Generation of Music for Inducing Emotive Response” and “User-Controlling Expressed Emotions in Music with EDME”) and the green to the generation of harmonic progression and jazz.

The clustering with the highest silhouette score for 2010 was $k = 11$ (see silhouette score comparison in Figure 2). In Figure 7, we can see that since the corpus is small, the clusters contain very few documents, but the several topics (that will appear also in the expanded datasets with the consecutive years) are announced, such as *lexical creativity-story generation* (keywords: story, knight, narrative, jaguar), *reasoning/association/bisociation* (papers: “Constructing Conceptual Spaces for Novel Associations”, “Bisociative Knowledge Discovery”. . . , “Domain Bridging Associations” and “Some Aspects of Analogical Reasoning in Mathematical Creativity”). The largest cluster refers to *visual creativity*, with the keywords: image, filter, darcus, robot, collage, fractal, but also several keywords due to noisy clustering—e.g., chat. (The term *darcus* is a lemmatised version of the system DARCI (mistakenly but reasonably assigning to the term a Latin origin)). The papers comprise “The Painting Fool”, “Swarm Painting Atelier”, “A Fractal Approach Towards Visual Analogy”.

Over the years, the clustering becomes more interesting, since we have more documents. So, for instance, in the vi-

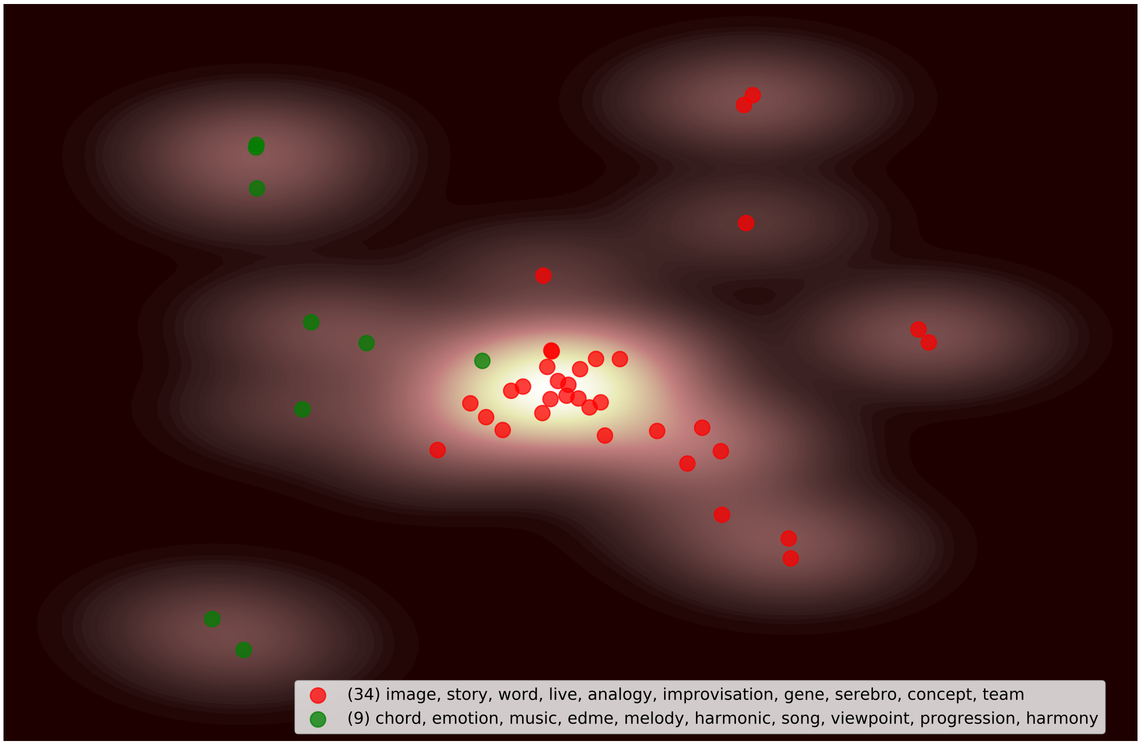

Figure 6: ICCC 2010 proceedings papers clustered into 2 clusters.

sualisation of papers from 2010–2015 in Figure 8, the CC domains are very clearly separated and characterised by corresponding keywords. *Music* is visible in yellow, *stories* and *games* are corresponding to the blue and green, respectively, in red we have *poetry* (keywords: poem, poetry, flowchart, syllable, rhyme, word, bengali, flowr), while in purple are other documents, including a clear coverage of *image*.

Unsurprisingly, the most interesting is the topic clustering on the entire corpus. The highest silhouette scores were returned for k=10 and k=11. We first analyse the k=10 clustering results. Since it covers the ICCC conceptualisation with the entire paper selection, we describe it in more detail, in terms of our category name and the associated keywords:

1. **Poetry** poem, poetry, flowchart, rhyme, syllable, word, bengali, expert, tweet, grammar, simile, template, constraint, text, poetryme
2. **Games** game, angelina, player, mechanic, utterance, jam, miner, mechanics, gameplay, rogue, spaceship, play, suspect, agent, designer
3. **Concepts** blend, icon, i1, blender, amalgam, conceptual, ontology, i2, colimit, space, optimality, goguen, workflow, input, relation
4. **Music** musical, music, chord, improvisation, musebot, melody, accompaniment, pitch, jazz, lyric, composition, musician, harmonic, song, participant
5. **Story and Narrative** story, character, narrative, knight, jaguar, action, plot, mexica, tension, enemy, princess, event, storyteller, scene, rez
6. **Image** darcus, image, adjective, synset, rendering, painting, fool, artifact, pareidolia, icon, volunteer, fiery, association, peaceful, train
7. **Embodiment and Choreography** robot, dancer, movement, choreographer, dance, empowerment, choreography, motion, agent, robotic, embody, antagonistic, sensor, choreographic, keyframe
8. **Cuisine** artifact, recipe, ingredient, surprise, novelty, rdc, haiku, card, apparel, artefact, cocktail, expectation, maze, inspiring, regression
9. **Conceptualising CC** cc, id, mlcc, copula, additive, artifact, preference, attribute, iccc, gaver, ig, function, student, marginal, intentional
10. **Other (not classified)** image, agent, object, node, association, analogy, word, metaphor, shape, painting, concept, conceptual, software, fitness, fig

As can be seen from the keywords, some clusters are pure while others include noise. We have yet to perform an extensive cluster evaluation, but we provide in Table 1 a full list of documents for the first topic cluster, where the precision is very high. The papers are available on the web[5], so minimal references are given here.

In addition, we compared the results of k=10 to k=11, to see if the clustering results are stable. We have observed that the *Poetry* cluster remains exactly the same (contains the same papers). The same holds true for the cluster (*Games*),

[5] http://computationalcreativity.net

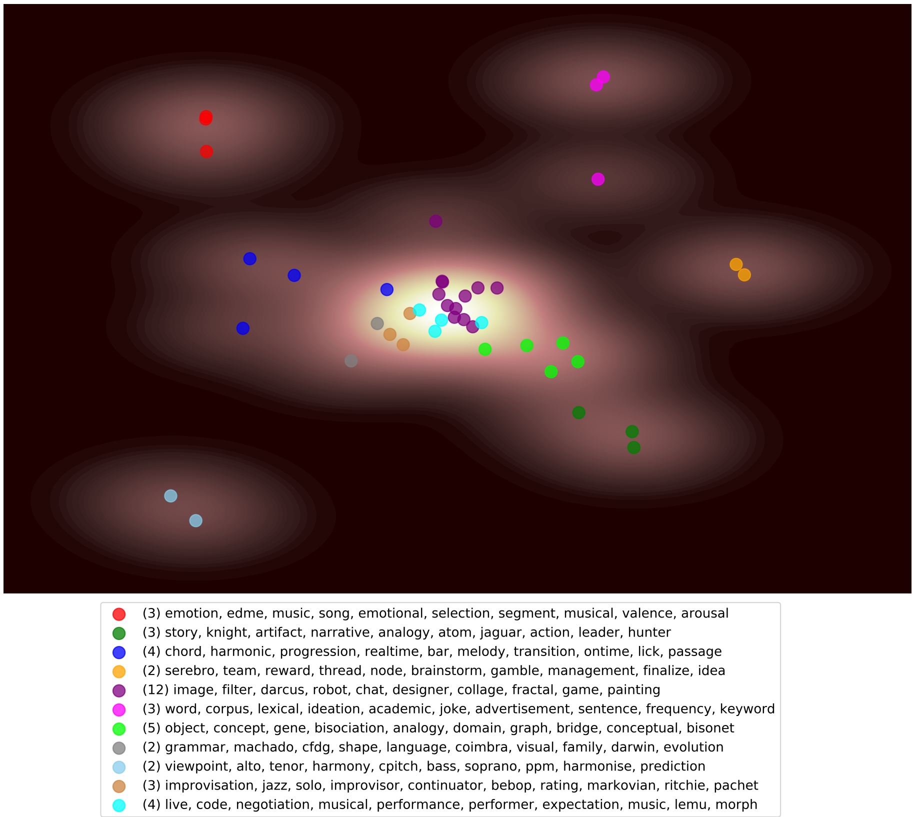

Figure 7: ICCC 2010 proceedings papers clustered into 11 clusters.

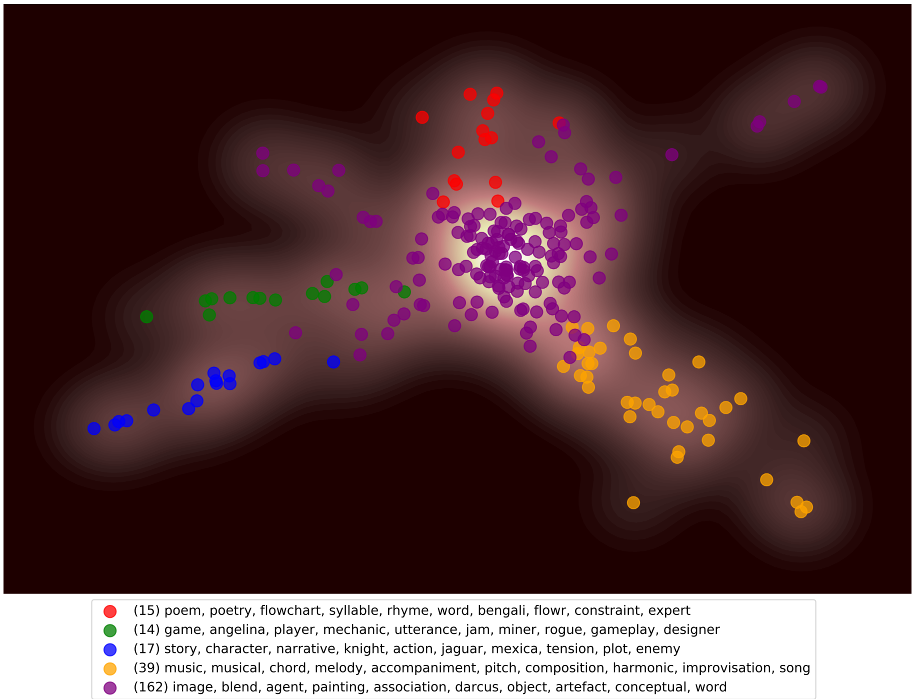

Figure 8: Papers of 2010–2015 ICCC proceedings clustered into 5 clusters.

the cluster related to conceptual blending (cluster *Concepts*), *Story and Narratives* and *Embodiment and Choreography*. Also the *Cuisine* cluster remains the same, but covers some papers, which do not belong to this cluster. For instance, the paper by Dan Ventura "Mere Generation: Essential Barometer or Dated Concept?" questions general principles of creativity, where recipes are just one of the several examples that are used in discussion. In Visual creativity (cluster Images), the only difference is in an additional paper added to the cluster with k=11, which is the paper describing the event You Cant Know my Mind. The cluster *Conceptualising CC*, contains one more paper in the cluster of k=10, which is in k=11 unclassified: this is our CC conceptualisation attempt (Pollak et al., 2016), for which it is understandable that it is not fixed to a single cluster, since it discusses different topics of computational creativity. The biggest difference can be observed in the *Musical* cluster, which is in the setting with k=11 split into two distinct clusters, with the following keywords:

- Music-C1: *musebot, musical, agency, improvisation, musician, music, jazz, participant, interaction, ensemble, performer, bown, improvise, kelly, practice*
- Music-C2: *music, chord, musical, melody, accompaniment, lyric, pitch, harmonic, composition, song, audio, markov, edme, beat, bass*

The conceptualisation across the years provides the clustering where the papers in each cluster can be used as reading material for the new members joining the ICCC community and being especially interested in a specific subdomain.

Conclusions and Future Work

This paper presents an overview of ICCC proceedings topics, achieved by the proposed methodology composed of data preprocessing, clustering and cluster visualisation. Since computational creativity is still a relatively new research field, it is still possible for the researcher to have an overview of the field as a whole, but with the growth of the field this will no longer be possible. Therefore, it is useful to provide a transparent and accessible overview of topics and categorised papers for sub-domains. We consider that this is very important especially for the incomers to the field.

We presented the results of analyzing different document sets and found out that the clustering results are mostly meaningful, allowing the expert to easily recognise the topics (e.g., musical creativity, story generation, poetry generation, visual creativity, culinary creativity, conceptual creativity, etc.). We experimented with automated discovery of the optimal number of cluster using the silhouette score but so far the results were not conclusive, since they did not fully align with human observations using 2D visual representations.

We will continue to work towards the automation of the process including clustering, concept naming, tracking topic changes within the selected domains, and computationally creating correct narratives over the history of computational creativity.

Table 1: ICCC papers captured in Cluster 1: Poetry

Bay, B., Bodily, P., and Ventura, D. (2017). Text transformation via constraints and word embedding.

Charnley, J., Colton, S., and Llano, M. T. (2014). The FloWr Framework: Automated Flowchart Construction, Optimisation, Alteration for Creative Systems.

Colton, S. and Charnley, J. (2013). Towards a flowcharting system for automated process invention.

Colton, S., Goodwin, J., and Veale, T. (2012). Full-FACE poetry generation.

Corneli, J., Jordanous, A., Shepperd, R., Llano, M. T., Misztal, J., Colton, S., and Guckelsberger, C. (2015). Computational poetry workshop: Making sense of work in progress.

Das, A. and Gambäck, B. (2014). Poetic Machine: Computational Creativity for Automatic Poetry Generation in Bengali.

Gervás, P. (2011). Dynamic inspiring sets for sustained novelty in poetry generation.

Gross, O., Toivanen, J. M., Lääne, S., and Toivonen, H. (2014). Arts, News, Poetry — The Art of Framing.

Kantosalo, A., Toivanen, J. M., and Toivonen, H. (2015). Interaction evaluation for human-computer co-creativity: A case study.

Lamb, C., Brown, D. G., and Clarke, C. (2015). Human competence in creativity evaluation.

Lamb, C., Brown, D. G., and Clarke, C. L. (2017). Incorporating novelty, meaning, reaction and craft into computational poetry: a negative experimental result.

Lamb, C., Brown, D. G., and Clarke, C. L. A. (2016). Evaluating digital poetry: Insights from the cat.

Misztal, J. and Indurkhya, B. (2014). Poetry generation system with an emotional personality.

Oliveira, H. G., Hervás, R., D'ıaz, A., and Gervás, P. (2014). Adapting a Generic Platform for Poetry Generation to Produce Spanish Poems.

Oliveira1, H. G. and Alves, A. O. (2016). Poetry from concept maps – yet another adaptation of poetryme's flexible architecture.

Rashel, F. and Manurung, R. (2014). Pemuisi: a constraint satisfaction-based generator of topical Indonesian poetry.

Tobing, B. C. and Manurung, R. (2015). A chart generation system for topical metrical poetry.

Toivanen, J. M., Järvisalo, M., and Toivonen, H. (2013). Harnessing constraint programming for poetry composition.

Toivanen, J. M., Toivonen, H., Valitutti, A., and Gross, O. (2012). Corpus-Based generation of content and form in poetry.

Acknowledgements

The work of VP, NL and SP was partly funded by the research programme Knowledge Technologies, funded by the Slovenian Research Agency. GW very gratefully acknowledges the financial and academic support of the Jožef Stefan Institute, during his sabbatical there in October and November 2017.

References

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.

Bradford, R. B. 2008. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, 153–162. New York, NY, USA: ACM.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence*, ECAI'12, 21–26. Amsterdam, The Netherlands, The Netherlands: IOS Press.

Deerwester, S. 1988. Improving Information Retrieval with Latent Semantic Indexing. In Borgman, C. L., and Pai, E. Y. H., eds., *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, volume 25. Atlanta, Georgia: American Society for Information Science.

Ester, M.; Kriegel, H.-P.; Sander, J.; and Xu, X. 1996. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, 226–231. AAAI Press.

Fortuna, B.; Mladenič, D.; and Grobelnik, M. 2006. Semi-automatic construction of topic ontologies. In *Semantics, Web and Mining: Joint International Workshops, EWMF 2005 and KDO 2005, Revised Selected Papers*, 121–131. Springer.

Gruber, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2):199–220.

Juršič, M.; Mozetič, I.; Erjavec, T.; and Lavrač, N. 2010. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *J. univers. comput. sci.* 16:1190–1214.

Kiss, T., and Strunk, J. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics* 32(4):485–525.

Loughran, R., and O'Neill, M. 2017. Application domains considered in computational creativity. In *Proceedings of ICCC 2017*. Association for Computational Creativity.

Marcus, M.; Kim, G.; Marcinkiewicz, M. A.; MacIntyre, R.; Bies, A.; Ferguson, M.; Katz, K.; and Schasberger, B. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, 114–119. Stroudsburg, PA, USA: Association for Computational Linguistics.

Pollak, S.; Boshkoska, B. M.; Miljkovic, D.; Wiggins, G.; and Lavrač, N. 2016. Computational creativity conceptualisation grounded on iccc papers. In François Pachet, Amilcar Cardoso, V. C. F. a. G., ed., *Proceedings of ICCC 2016*, 123–130. Association for Computaitonal Creativity.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20:53 – 65.

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

Smith, B. 2003. Chapter 11: Ontology. In Floridi, L., ed., *Blackwell Guide to the Philosophy of Computing and Information*, volume 7250. Blackwell. 155–166.

Tenenbaum, J. B.; Silva, V. d.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323.

Webber, W.; Moffat, A.; and Zobel, J. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* 28(4):20:1–20:38.