Augmenting the Evaluation and Mapping of Progress in Scientific Research – A Human-Machine Symbiosis Perspective

Andrej Dobrkovic[1](✉), Daniel A. Döppner[2], Maria-Eugenia Iacob[1],
and Jos van Hillegersberg[1]

[1] University of Twente, Enschede, Netherlands
{a.dobrkovic,m.e.iacob,j.vanhillegersberg}@utwente.nl
[2] University of Cologne, Cologne, Germany
doeppner@wim.uni-koeln.de

Abstract. In this paper we propose and demonstrate a software tool for symbiotic human-machine analysis, applicable for structured literature reviews (SLR). We present a seed-based search of bibliographic information, resulting in document clustering and graph visualization. Through a collaborative human-machine effort we show how to detect potential bridging articles and paradigm shifts. The overarching goal is to support the SLR process, especially for developing fields of science, as well as interdisciplinary fields, where similar concepts can be overlooked as they are associated with different keywords and belong to different groups, yet share common ideas. Finally, we demonstrate the application of the tool with two literature search and visualization examples.

Keywords: Human-machine symbiosis · Structured literature review
Data mining · Augmentation

1 Introduction

A typical structured literature review is a time-consuming process. In this paper we present and demonstrate a tool to augment the abilities of researchers to perform SLR, by enabling them to quickly identify potential areas of interest and the most important publications belonging to them. We base our approach on the idea of symbiotic interaction between human and machine [1] to leverage the processing power of computers to visualize and calculate measures in document networks, in combination with humans' creativity and visual perception for reasoning.

By adopting a "seed-based approach", as alternative to traditional keyword-based search, we rely on the human to identify the starting point for the search and select the most influential articles (i.e., the seeds) in the investigated domain. Using citations and references of the selected articles, the machine identifies related documents and connections between them. The results are transformed into a directed graph. Data mining is then applied to cluster the documents and provide visual cues to the articles based on the network analysis. By doing so, the human researchers can quickly analyze large

W. Karwowski and T. Ahram (eds.), *Intelligent Human Systems Integration*,
Advances in Intelligent Systems and Computing 722,
https://doi.org/10.1007/978-3-319-73888-8_56

volumes of scientific papers, understand their relations and dependencies, and identify potentially high-impact publications for further investigation. In the following, we describe the underlying approach, provide details about the tool implementation and demonstrate its application.

2 Solution Design

2.1 Core Idea

The core idea behind this approach is to base knowledge extraction and discovery around the seeds; papers, deemed to be most relevant to the process by the researcher, and expand the search from there. To help the researcher in situations when the searches return a vast amount of potentially relevant documents, we want to be able to provide an additional search procedure for quick identification and zoom in on the most important articles. This is accomplished by transforming the search results into a directed graph, in which every document becomes a node, and edges are citations and references. Assuming that highly influential documents have a number of edges greatly exceeding that of less important ones, the tool provides visual cues to the researcher. Since the qualitative evaluation of the content of the scientific material is beyond the capabilities of the machine, the human is required to scan the suggested publications and decide on their relevance. If there is a need to further expand the search, the process can be iteratively repeated by adding the most important of the newly discovered documents as additional seeds. The process is illustrated in Fig. 1 and contains the following steps: (1) selection of the seed(s) and search criteria, (2) knowledge extraction from bibliographic databases, such as Google Scholar and Scopus, (3) data processing – results transformation into a graph structure and clustering, and (4) data visualization.

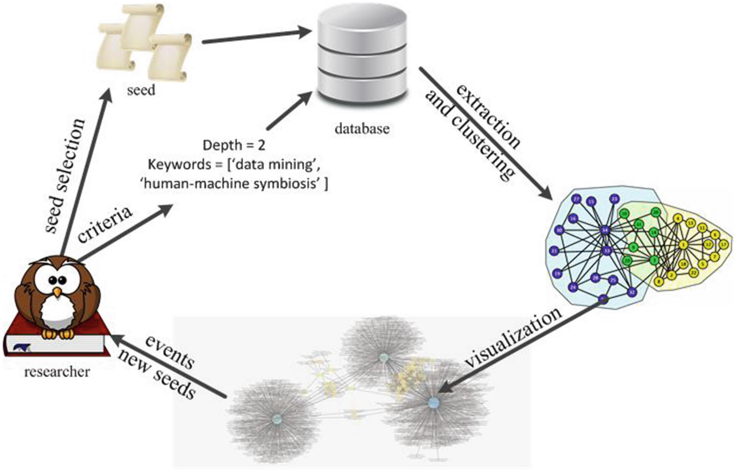
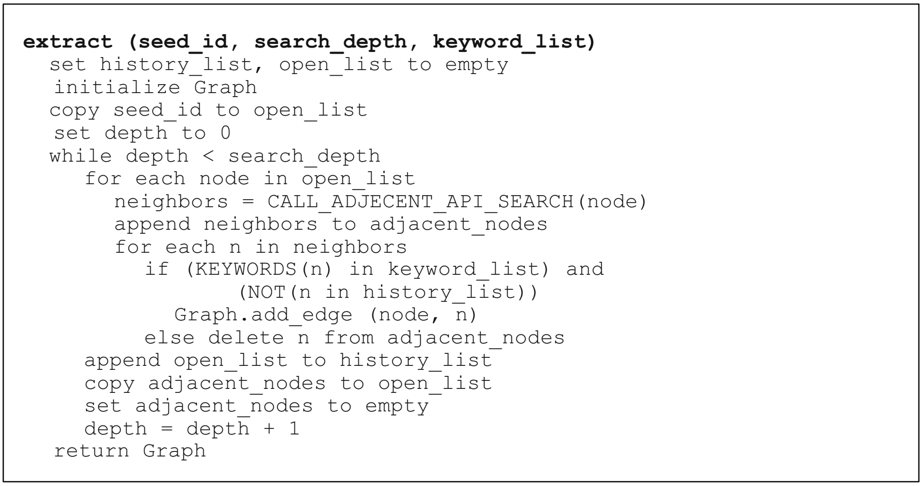


Fig. 1. The process of generating citation network

2.2 Extraction Algorithm

The extraction is based on a symbiotic human-computer approach. Human directs the search by deciding on the seeds to be used, keyword filters, and search depth. The machine relies on brute force to do the extraction and classification.

The extraction algorithm uses database specific APIs to connect to the bibliographic database. Then, the search function is called to perform forward and backward search starting from the given seed's unique ID. The forward search includes all documents that cite the initial document and backward search includes all referenced documents. For each retrieved document, the process is repeated according to the given search depth parameter. Optionally, a set of keywords can be given, limiting the algorithm to retrieve only the documents that contain at least one of the given keywords, thus preventing the search from expanding into non-related scientific fields. The pseudo-code for extraction and graph generation is shown in Fig. 2.



```
extract (seed_id, search_depth, keyword_list)
  set history_list, open_list to empty
  initialize Graph
  copy seed_id to open_list
  set depth to 0
  while depth < search_depth
     for each node in open_list
       neighbors = CALL_ADJECENT_API_SEARCH(node)
       append neighbors to adjacent_nodes
       for each n in neighbors
          if (KEYWORDS(n) in keyword_list) and
                 (NOT(n in history_list))
            Graph.add_edge (node, n)
          else delete n from adjacent_nodes
     append open_list to history_list
     copy adjacent_nodes to open_list
     set adjacent_nodes to empty
     depth = depth + 1
  return Graph
```

Fig. 2. Pseudo-code for document extraction and graph generation

2.3 Visualization

The grouping of nodes is based on community extraction from large networks. We apply Louvain Modularity [2] for the clustering. To improve visibility and reduce edge overlap, we use force directed graph layout. In our test cases, we achieve the best visual results with node coordinates generated by ForceAtlas2 [3]. Additionally, we use visual styling, to help the user quickly identify essential information: (1) node size – number of edges, (2) node color – modularity/cluster, (3) node shape – highlighting potentially important documents, and (4) edge color – publication year of the related document.

The symbiotic component of the approach is expressed in the following division of tasks. The machine plots the data in a manner suitable for the human to quickly browse through large number of documents. The human is responsible for the qualitative evaluation of the relevance of the documents clusters.

A visualization example is given in Fig. 3. We used a single seed and limited the search depth to two. The extraction algorithm finds 669 related documents, which are grouped into four different clusters (see node colors, i.e. red, blue, green, and yellow in Fig. 3). The number of citations determines the size of each node. As it can be seen, cluster centers are larger than other nodes. To enable the researcher to quickly observe the field's development, edge color indicates the publication year of the connected document. As the tool should raise the researcher's awareness to potentially important papers, these are rendered as squares, while the remaining nodes are rendered as circles.

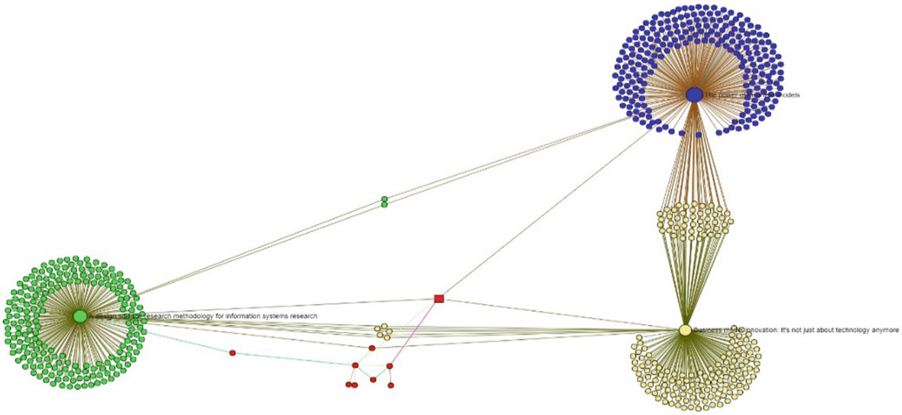

The power of business models
A design science research methodology for information systems research
Business model innovation: It's not just about technology anymore

Fig. 3. Visualization example

2.4 Events

Even a single seed search with low depth has the potential to return a large number of results. Although, visualization with clustering helps the researcher to explore the field, the tool should also assist by highlighting potentially significant documents. We refer to these documents as "events" and they represent one of the following: (i) paradigm shifts or unification events, and (ii) bridging articles. According to Moody et al. [4] and Kuhn [5], a unification event is the moment when the consensus is achieved in a specific scientific field, while paradigm shifts are the result of scientific revolutions that shift the scientific field. We use the term bridging article for the documents that span different fields of science and as such provide excellent candidates for future seeds.

A formal mathematical specification of these events is not possible, yet potential events can be detected based on their connectivity with other clusters. Relying on human-machine symbiotic approach, we use the human to define a certain threshold t, and then let the machine search for all nodes with a number of edges e exceeding it:

$$e > t \tag{1}$$

3 Demonstration

We demonstrate the concept using two examples. For the first example, we aim for a small graph to reduce edge overlap and improve visual clarity. Therefore, we pick a relatively recent article with limited amount of citations and references as seed. Input parameters and output results are given in Table 1.

Table 1. Parameters for examples 1 and 2

Example 1		
Input	No. seeds	1
	Seed paper	Creating a business case from a business model
	Keyword filter	Business case, Business model, Method, Methodology
	Database	Scopus
	t	3
	Date executed	3-jul-17
Output	No. retrieved documents	670
	No. clusters	4
	No. events	1 (the seed itself)
	Graph	Figure 3
Example 2		
Input	No. seeds	2
	Seed paper	Man-computer symbiosis
		Man versus Machine or Man + Machine?
	Keyword filter	Intelligence amplification, Intelligence augmentation, Human-machine symbiosis, Human-computer symbiosis, Human-machine collaboration, Human-machine cooperation
	Database	Scopus
	t	3
	Date executed	14-sep-17
Output	No. retrieved documents	1244
	No. clusters	30
	No. events	4
	Graph	Figure 4

For the given seed, “Creating a business case from a business model” by Meertens et al. [6], the extraction algorithm retrieves 670 documents (including the seed) and groups them in four different clusters. From the graph in Fig. 3, we can identify clusters about design science methodology, business modeling, and business model innovation. Since the event threshold *t* was set to 3, only the seed itself is visualized as bridging paper. By comparing references, and citations from the seed and the output graph, no aberrations can be identified, therefore we conclude that the visualized components provide acceptable results.

The second example combines two seeds. We use “Man-computer symbiosis” by Licklider [1] as one of the most influential papers in this domain, together with one relatively new publication sharing the similar idea, “Man versus Machine or Man + Machine?” by Cummings [7]. The search retrieves 1244 documents and performs the clustering as shown in Fig. 4. As expected, Licklider’s paper is assigned the central location in the graph and identified as potential event. We detect 30 clusters in total and the tool draws attention to three additional event candidates. Two are centers of new clusters related to human computation and human interaction with automation. Also, due to high connectivity, the paper “Exploring Design Principles for Human-Machine Symbiosis: Insights from Constructing an Air Transportation Logistics Artifact” by Döppner et al. [8] is signaled out to the human as potential bridging event.

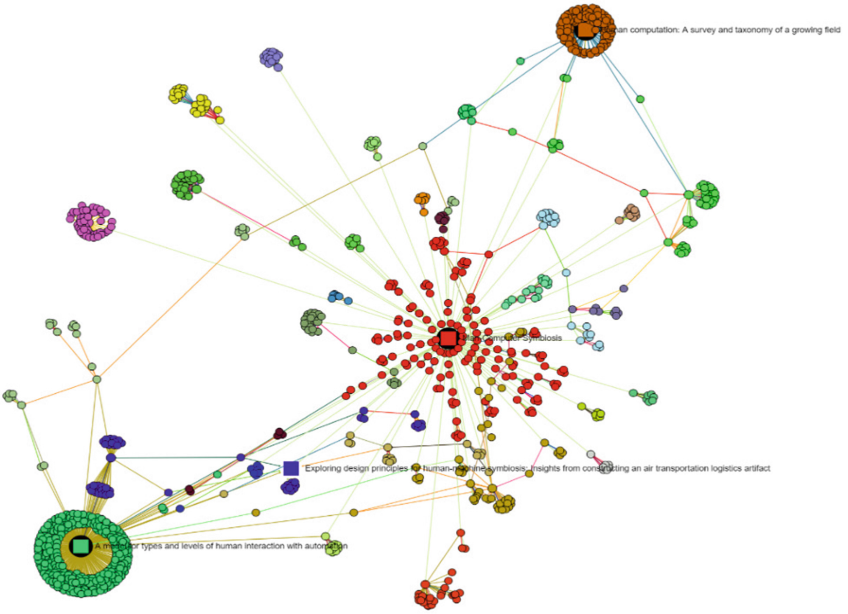

Human computation: A survey and taxonomy of a growing field
Man-Computer Symbiosis
Exploring design principles for human-machine symbiosis: Insights from constructing an air transportation logistics artifact
A model for types and levels of human interaction with automation

Fig. 4. Graph representation of a two-seeds search

In this paper, we have presented our tool to extract, cluster, and visualize relationships between publications. It enables researchers to explore large collections of scientific articles and identify potentially areas of interest for structured literature review. Based on the tests we performed, we expect the tool to be effective for doing structured literature reviews. We expect the tool to be most beneficial in developing fields of science, as well as interdisciplinary fields, where similar concepts can be labeled with different keywords, yet share common elements. Through visualization and signaling event documents, we seek to provide a rigorous search tool for any field of science.

References

1. Licklider, J.C.: Man-computer symbiosis. IRE Trans. Hum. Factors Electron. **1**, 4–11 (1960)
2. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. **2008**(10), P10008 (2008)
3. Jacomy, M., Heymann, S., Venturini, T., Bastian, M.: Forceatlas2, a continuous graph layout algorithm for handy network visualization. Medialab center of research 560 (2011)
4. Moody, D.L., Iacob, M.-E., Amrit, C.: In search of paradigms: identifying the theoretical foundations of the IS field (2010)
5. Kuhn, T.S.: The route to normal science. In: The Structure of Scientific Revolutions, vol. 2, pp. 10–22 (1970)
6. Meertens, L.O., Starreveld, E., Iacob, M.-E., Nieuwenhuis, B.: Creating a business case from a business model. In: International Symposium on Business Modeling and Software Design 2013, pp. 46–63. Springer, Heidelberg (2014)
7. Cummings, M.M.: Man versus machine or man + machine? IEEE Intell. Syst. **29**(5), 62–69 (2014)
8. Döppner, D.A., Gregory, R.W., Schoder, D., Siejka, H.: Exploring Design Principles for Human-Machine Symbiosis: Insights from Constructing an Air Transportation Logistics Artifact (2016)