INTERNATIONAL DESIGN CONFERENCE - DESIGN 2016
Dubrovnik - Croatia, May 16 - 19, 2016.

A DATA MINING APPROACH TO ASSIST DESIGN KNOWLEDGE RETRIEVAL BASED ON KEYWORD ASSOCIATIONS

F. Shi, J. Han and P. R. N. Childs

Keywords: design knowledge, problem solving, data mining, keywords association, association rules

1. Introduction

In the engineering design process, creative idea generation (ideation) is a key activity in design thinking and engineering thinking. It can be viewed as a problem-solving activity which produces ideas and solutions for all stages from conceptual mechanism design to detailed structure design. Even though a lot of creativity tools have been developed to facilitate creative idea generation [Chakrabarti and Bligh 1996], [Altshuller 1997], [Childs 2014], idea generation still depends highly on an important prerequisite: design knowledge retrieval, which involves searching and organizing professional knowledge for design and engineering queries. Previous research studies investigating novice engineers who had one or two years of experience, found that they usually were not aware of what information they needed [Ahmed and Wallace 2003]. Design knowledge retrieval needs to satisfy the knowledge requirements and queries of engineers by providing effective knowledge and relevant expertise. Thus, robust knowledge retrieval approaches are key for creative ideas generation in the design process.
Aurisicchio [2005] found that interaction with colleagues was the main method for knowledge retrieval in the industrial field. From total queries of engineers, 70% were sourced through interaction with colleagues, and the rest 30% were sourced by documents, drawings and databases. From the total queries addressed to colleagues, 65% were sourced through face to face interactions, and 35% were sourced through email and phone. However, interaction with colleagues is not always very efficient and effective, which involves subjective bias and highly depends on the nature of queries. Ahmed [2006] proposed an approach based on indexed documents to assist designers with their queries and designs. This approach assigned four properties to each document according to products, design processes, functions and issues. Each property of the documents was indexed by a knowledge management system, where a network built upon the indices can help prompt design tasks.

2. State of art

Recently, computing and data mining have been widely applied to gain fundamental insights and knowledge from massive data [Zaki and Meira Jr. 2014], which enables the growth of semantic technology and ontology-based knowledge representation. Tran and Cimiano [2007] presented an approach for translating user keywords queries into database language queries to explore and retrieve knowledge from an existing database by semantics-based searching, however the method to build an appropriate semantic database was not mentioned. Text comparison and structuring have been developed to recognize the content similarities between text items for relevant information retrieval and documents searching [Salton and Buckley 1991], [Salton et al. 1996], [Salton et al. 1997]. Most of these

tools are based on purely statistical techniques, while the inherent semantic relations between knowledge concepts are rarely explored.

Some researchers implemented WordNet as a semantic base to make associations between words in order to help people with associative thinking [Budanitsky and Hirst 2001], [Pedersen et al. 2004], [Seco et al. 2004]. WordNet can be viewed as a set of trees constructed based on hypernym and hyponym. Figure 1 shows an example of two knowledge trees in WordNet. Words can only be efficiently retrieved between hypernym(parent) and hyponym(child) within the same branch of one tree, but there is no effective way to associate two words located in two separate trees, e.g. "hydrophobic membrane" and "desalination". From design problem-solving perspective, "hydrophobic membrane" is actually a creative solution for "desalination" in chemical engineering. Therefore, WordNet lacks the capacity to make this kind of creative association for design knowledge retrieval.

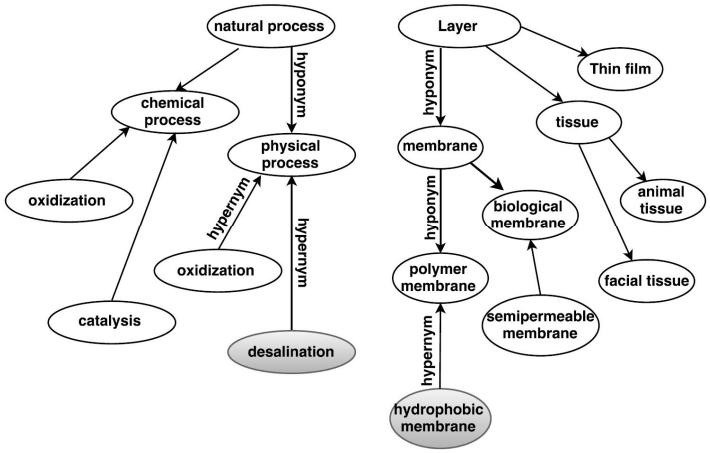

Figure 1. Two separate trees data in WordNet

ConceptNet [Liu and Singh 2004], [Speer and Havasi 2013], is another relational knowledge network constructed through a pattern recognition process by data mining technique. ConceptNet extracts the predetermined patterns in text mining and then represents the knowledge information as Graph data including edges and vertices. Vertices represent the objects while edges stand for the relations between them. For example, Figure 2 illustrates the partial view of ConceptNet representing knowledge in a cluster of related concepts [Speer and Havasi 2012]. Superior to WordNet, ConceptNet achieves a higher level association scope involving relations of fundamental function, behavior and structure, which are relevant to design thinking. However, at this moment, ConceptNet is still a common sense network representing general relational knowledge, which is not capable to meet the professional knowledge and expertise needs of engineer queries.

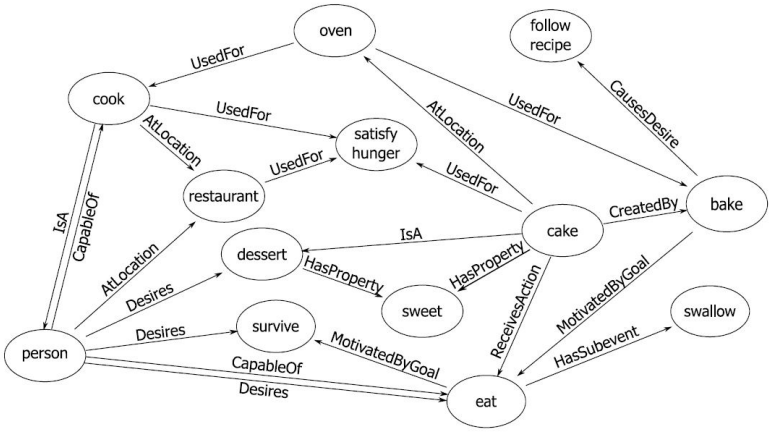

Figure 2. A local view of knowledge ConceptNet [Speer and Havasi 2012]

In this research, an approach to develop a professional knowledge and expertise network was proposed using data mining techniques in order to assist design knowledge retrieval. Keywords of journal papers were used as the mining resources since they represent consice and specialized knowledge concepts and have inherent high-level associations from design and engineering perspectives. An algorithm was developed to generate association rules within the keywords. A web crawler was implemented with Elsevier API for keyword extraction, and the knowledge network constructed was saved into a database system using a graph data structure. Finally, software with a simple Graphic User Interface (GUI) has been developed for users to make associations within this network for knowledge retrieval, and a design case study was carried out to evaluate the effectiveness.

3. Research approach

According to the above discussion, two gaps were found in applying current semantic networks into practical design knowledge retrieval. The research goals of this paper are deduced to fill these gaps:

- The relations in WordNet are basically constructed on hyponym and hypernym. This cannot satisfy design knowledge retrieval which involves high level association processes based on relations about methods, functions, materials, strucutres, etc. Thus, our research aims to explore the inherent associations between knowledge concepts from design and engineering perspectives.
- ConceptNet has well-established semantic relations between conceptual items, however, its knowledge domains are currently restricted in common sense, which are not suitable for responding to highly specialized engineering queries containing extensive expertise. The approach in this research will focus on building associations between professional knowledge concepts instead of common sense.

3.1 Knowledge form and source

The form and source of fundamental knowledge concepts are important to build the professional knowledge network. Concise forms or patterns are effective for representing knowledge concepts in order to produce brief and clear knowledge network, which is easy and efficient for users to retrieve relevant information. Therefore, in this research, technical keywords are applied to represent the knowledge concepts because of its simplicity and specificity.

Based on the two goals, in order to concentrate on professional knowledge rather than common sense, the keywords of large amounts of academic journal papers in engineering fields are applied as the knowledge source here, because the keywords of these journal papers tend to be highly specialized and cover a broad range of concepts involving methods, principles, processes, materials, functions and mechanisms in various engineering fields. Keywords of one individual academic paper usually show the research objects, research methods and even principles and materials discussed in the paper. Thus, the inherent correlations between research objects, research methods and principles studied in a journal paper would result in the inherent associations between the keywords of this paper. Therefore, the keywords of a single academic paper are actually automatically related to each other within a high-level academic association scope.

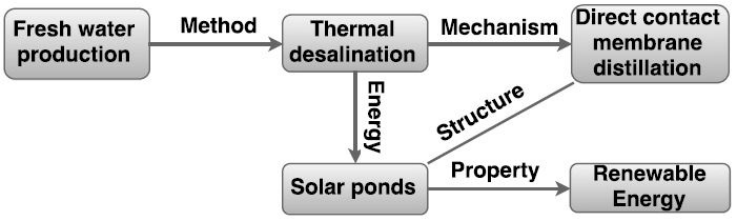

Figure 3. Example association between five keywords of an academic paper

Here is an example of one academic paper [Suárez et al. 2015] containing five sets of keywords: "Fresh water production", "Thermal desalination", "Direct contact membrane distillation", "Renewable

energy" and "Solar ponds". It can be seen that there exist high-level associations between these five sets of keywords from design engineering perspective, as shown in Figure 3 as a relational network. Similarly, every paper has its own keywords network. The networks of different papers can be combined together by using their common keywords as the joints. Figure 4 shows the combined keywords network by linking Figure 3 with a second paper [Rastegarpanah and Mortaheb 2016]. The keywords of the second paper are: "Direct contact membrane distillation", "Hydrophobic membrane", "Polyethersulfone", "Tetraethylorthosilicate", "Trimethylchlorosilane", in which the common keyword "Direct contact membrane distillation" of both papers is used as the joint to link the two networks together. The keywords of the second paper enrich the knowledge about material specifications for "Direct contact membrane distillation".

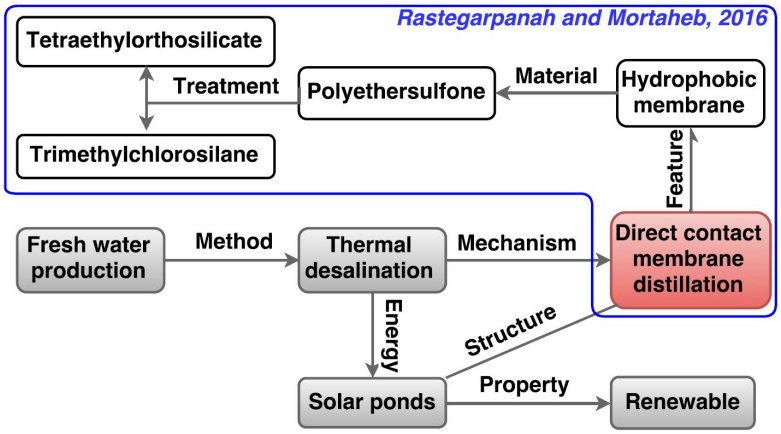

Figure 4. Combination of keywords network of two papers

In this way, a knowledge network can be constructed using the associations of keywords in millions of academic papers. This huge network representing specialized knowledge concepts with inherent high-level associations, has potential to be more powerful than WordNet and ConceptNet for design knowledge retrieval, and satisfies the two key goals above.

3.2 Data mining on association rules

However, for millions of papers in practice, it is impossible to manually construct the keywords network of each paper and combine millions of networks into a huge one by searching their common words as the join points, which would be extremely time-consuming. We can, however, use computational resources with data mining techology to realize this process. Data mining comprises powerful algorithms dedicated to massive data analysis, in which the itemset mining technique [Hastie et al. 2005], [Kantardzic 2014] can help exploring the associations between keywords to construct the knowledge network.

As illustrated in section 3.1, keywords within one academic paper usually have high-level associations with each other, which is similar to the classic market basket analysis problem in itemset mining [Witten and Frank 2005]. In market basket analysis, items in one basket are considered to be associated between each other. Thus, we can view each paper as a supermarket basket and the keywords of the paper as the items in the basket. Furthermore, the strength of associations between keywords can be evaluated by the $support\ value$ in itemset mining algorithm.

Let $\mathcal{P} = \{p_1, p_2, \dots, p_n\}$ be a set of academic journal papers, where p_i denotes one single paper in the paper set $\mathcal{P}$. Let $\mathcal{K} = \{k_1, k_2, \dots, k_m\}$ be a total collection of the keywords of all papers in $\mathcal{P}$. A set $K \subseteq \mathcal{K}$ can be called a subset in $\mathcal{K}$. Thus, a basket transaction can be expressed as a tuple of the form $\langle p, K(p) \rangle$, where $p \in \mathcal{P}$ is an unique academic paper, and $K(p) \subseteq \mathcal{K}$ is the set of the keywords contained in p. A transaction database $\mathcal{D}$ can be built as a list of the basket transactions. Table 1 shows an example of basket transaction database $\mathcal{D}$. Here $\mathcal{P} = \{p_1, p_2, \dots, p_6\}$ is the set of papers and $\mathcal{K} = \{k_1, k_2, \dots, k_8\}$ is the set of all the keywords in $\mathcal{P}$.

Table 1. Basket transaction database

p	$K(p)$
p_1	$k_2\ k_4$
p_2	$k_1\ k_2\ k_5\ k_6$
p_3	$k_1\ k_5\ k_7$
p_4	$k_2\ k_8$
p_5	$k_1\ k_2\ k_5$
p_6	$k_3\ k_5\ k_7$

The transaction database $\mathcal{D}$ is actually a relation between the paper set $\mathcal{P}$ and keywords set $\mathcal{K}$. We say that paper $p \in \mathcal{P}$ contains keyword $k \in \mathcal{K}$ iff $(p,k) \in \mathcal{D}$ and $k \in K(p)$. A transaction database is essential in itemset mining as it contains all the fundamental information we need to evaluate the associations between keywords and construct the knowledge network.

An association rule is an expression $k_i \rightarrow k_j$, where k_i and k_j are two keywords that $k_i,\ k_j \in \mathcal{K}$. The $support\ value$ of the rule in a transaction database $\mathcal{D}$ is the number of transactions in which both k_i and k_j co-occur:

$$sup(k_i \rightarrow k_j) = |\{p|\langle p, K(p)\rangle \in \mathcal{D} \text{ and } k_i, k_i \in K(p)\}| \tag{1}$$

For example, in Table 1, for association rule $k_1 \rightarrow k_5$, we found that papers p_2, p_3 and p_5 contain both k_1 and k_5. Thus, the $support\ value$ of the association rule $sup(k_1 \rightarrow k_5) = 3$. In our case, association has no direction as we can find that $sup(k_5 \rightarrow k_1)$ is also 3. The matrix in Table 2 shows the $support\ values$ of associations between all keywords in transaction database $\mathcal{D}$. Each cell in the matrix represents the $support\ value$ of the association rule between the keywords in row label and column label. This matrix is symmetric due to $sup(a \rightarrow b) = sup(b \rightarrow a)$. The diagonal line of the matrix where row label and column label are the same keyword, illustrates the frequence of occurrence of this keyword in transaction database $\mathcal{D}$. Efficient algorithms have been developed to compute the $support\ values$ of association rules in itemset mining such as Apriori Algorithm, Eclat Algorithm and FPGrowth Algorithm [Zaki and Meira Jr. 2014], which are all applicable to existing databases. In our case, an algorithm was developed to establish the database and calculate the $support\ values$ synchronously, which will be described later in Section 4.

Table 2. Support values of all association rules in database $\mathcal{D}$

	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8
k_1	3	2	0	0	3	1	1	0
k_2	2	4	0	1	2	1	0	1
k_3	0	0	1	0	1	0	1	0
k_4	0	1	0	1	0	0	0	0
k_5	3	2	1	0	4	1	2	0
k_6	1	1	0	0	1	1	0	0
k_7	1	0	1	0	2	0	2	0
k_8	0	1	0	0	0	0	0	1

A keyword knowledge network could now be constructed by using the generated association rules with corresponding $support\ values$. In data mining analysis, the network can actually be represented as a graph data structure. A graph is a pair $G = (V, E)$, where V is a set of vertices, and $E \subseteq V \times V$ is a set of edges. In our case, keywords are the vertices in graph $V = \mathcal{K} = \{k_1, k_2, \ldots, k_m\}$, and edges represent the association rules between keywords. An edge can be constructed between two vertices if the $support\ value$ of association between this two keywords is nonzero. Thus an edge can be expressed as a tuple:

$$e_{ij} = \langle k_i, k_j, sup(k_i \rightarrow k_j) \rangle, \text{ subject to } sup(k_i \rightarrow k_j) > 0 \quad (2)$$

where k_i, k_j are two vertices, and the nonzero $support\ value\ sup(k_i \rightarrow k_j)$ is the label of the edge. If there exists an edge between two vertices, we say the two vertices are adjacent. The edge set E can be formed as a list of such tuples $E = \{e_1, e_2, \ldots, e_N\}$. Figure 5 shows the keywords network graph based on Table 2.

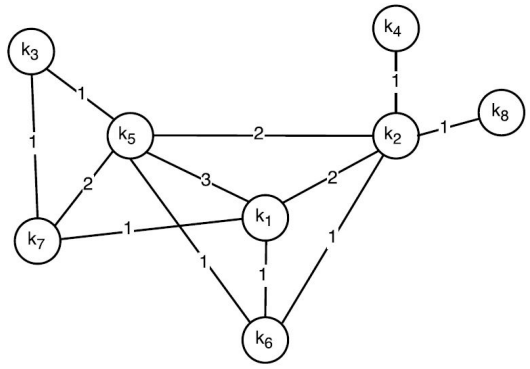

Figure 5. Graph representation of keywords network

3.3 Keywords retrieval process

In design knowledge retrieval, the core of engineering query can usually be summarized into a knowledge concept as a keyword. Design knowledge retrieval aims at searching and extracting relevant knowledge around this query. As explained in section 3.1, a keyword of a paper usually has inherent associations with the other keywords of this same paper, so a keyword in the graph has much stronger association with adjacent vertices rather than nonadjacent vertices. In addition, since the $support\ value$ between two keywords indicates the number of papers each of which contains both keywords, thus, larger $support\ value$ would illustrate stronger association. Therefore, the design knowledge retrieval process in our keywords network follows two principles simultaneously:

- Breadth first search (BFS) for adjacent vertices.
- Sort $support\ values$ in descending order for strong associations.

Take Figure 5 as an example. If the engineering query can be represented as k_1, the design knowledge retrieval process would follow BFS to extract adjacent k_5, k_2, k_6, and k_7 in sequence based on the descending order of $support\ value$. Then, each of these vertices will be orderly explored one after another in the same manner. Thus, k_3 will be extracted from k_5 before the extraction of k_4, k_8 from k_2. The retrieval process can proceed until there is no more vertex to explore. So, the final retrieval process would be: $k_1 \rightarrow k_5 \rightarrow k_2 \rightarrow k_6 \rightarrow k_7 \rightarrow k_3 \rightarrow k_4 \rightarrow k_8$.

4. Implementation

The practical implementation of the above theoretical approach involves the processes of data extraction, data construction and data interaction.

In this paper, a simple web crawler coupled with an Elsevier API [Dev.elsevier.com 2015] was programmed for keywords extraction from internet. The Elsevier API provides the access to journal papers included in ScienceDirect, and the web crawler extracted the keywords section of each paper accessed. A depth first search (DFS) method was conducted to crawl the corpus through ScienceDirect site map. The site map can be represented as a hierarchical tree as shown in Figure 6. In this research, we temporarily focus on the journal papers relevant to applied engineering fields as our samples to build the knowledge network. After running the crawler for 70 hours, more than 120,000 keywords had been extracted from about 76600 papers among 12 journals, which include «International Journal of Machine Tools and Manufacture», «International Journal of Mechanical Sciences», «Desalination», «Archives of Civil and Mechanical Engineering», «Applied Energy», «Chemical Engineering and Processsing», etc.

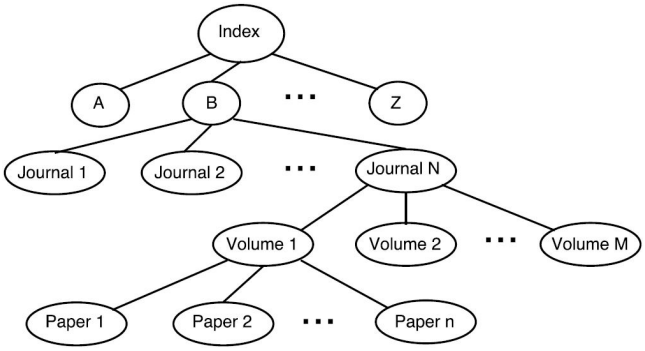

Figure 6. Hierarchy of ScienceDirect sitemap

Data construction was carried out simultaneously with data extraction. Whenever a paper was accessed, the keywords of this paper and their associations would be constructed and saved to a database which represented the keywords network. Since the network is a form of graph structure $G = (V, E)$, two relational tables are used to store the vertex set V and the edge set E respectively. The vertex table applies only one indexed column to store all the keywords $\mathcal{K}$, while the edge table contains three columns to respectively store two indexed vertices and one edge label in the form of $\langle k_i | k_j | sup(k_i \rightarrow k_j) \rangle$. The process of our algorithm to incorporate the keywords K of a single paper is described as follows: it firstly updates the vertex table through identifying and adding keywords which are new to the network. Then, it iterates to construct associations between every two words in set K. If the association already exists in the edge table, its $support\ value$ in the edge label will be increased, otherwise, the association will be saved as a new edge in the network. In our case, around 1,268,000 edges have been constructed in the database.

In data interaction, a simple piece of software implemented in Python has been developed for users to interact and retrieve with the keywords network, as shown in Figure 7. It internally follows the two retrieval principles discussed in section 3.3. Users can input the keyword of their query into the center textbox. By clicking "Associate mid" button, the most relevant knowledge concepts to the query word will be retrieved and displayed in the three surrounding textboxes. The remaining less related concepts can be obtained by repeatedly clicking the "Associate mid" button. During this process, associations can alternatively be generated from other keywords in surrounding textboxes by clicking "Associate top", "Associate left" or "Associate right". These buttons will place the keyword from surrounding textbox into the center textbox to make it as a new query, and proceed with the same retrieval process above. This improves the diversity and flexibility of associations in the network. Keywords already displayed will be automatically ignored when it is retrieved again, unless "Restart" button is clicked which clears the records of retrieval history. The "Back" button enables movement back to the previous association step, which helps flexibly controlling the retrieval activities in the keywords network.

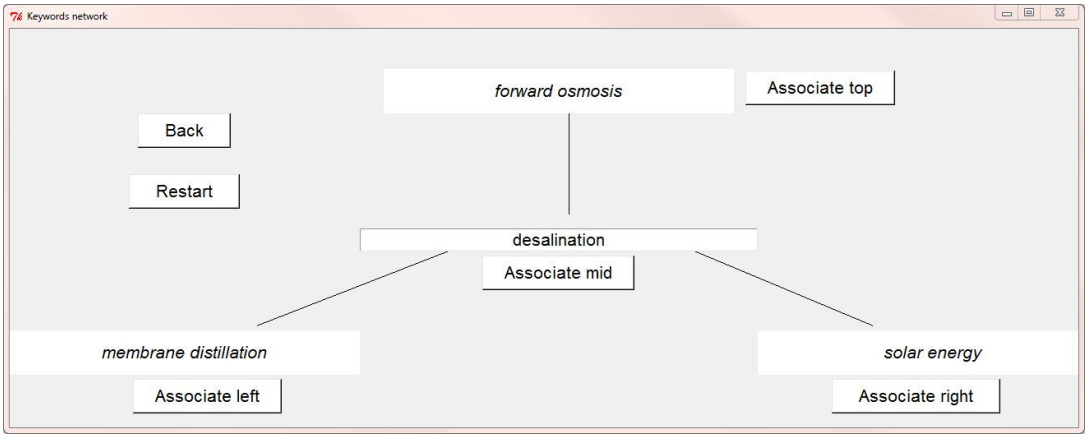

Figure 7. GUI for keywords network retrieval

5. Knowledge retrieval case study

This section focuses on the application of this software in a real world design knowledge retrieval process. The results and usage of the software can help justify the effectiveness of our proposed approach through a sustainable water purifier design case study.

The software is evaluated by using an engineering design project conducted in a design engineering school. The aim of this project is to develop a household product to convert seawater into drinking water for residential users near a coast. Since the software has explored knowledge about chemical engineering and purification technology, we can use the software to assist relevant knowledge concepts retrieval to generate ideas in this design process. In this project, "Desalination" which is able to represent the core of our engineering query, can be used as the start point of the retrieval process. Through this software, associations from "Desalination" have been generated to retrieve relevant knowledge concepts, from which additional associations were recursively retrieved to obtain further knowledge. As a result, the network built in this way is partly shown in Figure 8, in which we can see that there are much useful knowledge to support our design task. For example:

- For mechanism ideation in conceptual design, the software starts the retrieval process from "Desalination" as the core query. Many relevant knowledge concepts are retrieved from "Desalination" such as "reverse osmosis", "forward osmosis", "membrane distillation", "osmotic distillation", "concentration polarization", "multi-stage flashing (MSF)", "multi-effect distillation", "capacitive deionization", "electrodialysis", etc. All these technical concepts which are the widely-applied engineering solutions to desalination, can be used as the feasible mechanisms of our water purifier product.
- By further exploring any of the above mechanisms recursively, for example, from "membrane distillation", the software can link to "direct contact membrane distillation", "air gap membrane distillation", "sweep gas membrane distillation" and "vacuum membrane distillation", which are four specific types of "membrane distillation". This kind of information is very useful for the architecture of the product in embodiment design stage.
- Again, the terms "Hydrophobic membrane" along with its materials "polypropylene (PP)", "polyethersulfone (PES)", "polytetrafluoroethylene (PTFE)" and "polyvinylidene fluoride (PVDF)" can also be retrieved from "membrane distillation". "Hydrophobic membrane" is the key component in "membrane distillation", and "PP", "PTFE", "PES" and "PVDF" are all the commonly-used materials of "Hydrophobic membrane". Such information about structures, properties and materials could be helpful in the detailed design.
- On the other hand, "solar energy", "photovoltaic", "hybrid system" and "solar thermal collector" are also retrieved, which effectively provide feasible solutions for the energy supply of a product.
- Furthermore, the software also retrieves other relevant concepts such like "Biofouling". This can help remind designers to also consider other problems such as water treatment and purification in addition to the desalination function of the product. Based on "Biofouling", corresponding information such as "ozonation", "chlorination", "biofilm" and "EPS" are recursively retrieved.
- Finally, with above information, we are able to draft a design solution for consideration: based on membrane distillation, our sustainable water purifier could implement air-gap membrane distillation structure with PTFE hydrophobic membrane, use ozonation treatment method and is powered by solar thermal and photovoltaic hybrid energy system.

The associations generated by the software are shown in Figure 8. Each newly acquired word or concept is analyzed by expertise based on whether it's a piece of useful knowledge for our following engineering problems and design activities. We marked effective associations in blue color if they are meaningful for exploring relevant knowledge concepts to assist our design task, while useless and unhelpful associations were marked in red. The result in Figure 8 shows that about two thirds of the associations generated by the software are evaluated to be useful, as judged by the engineering team.

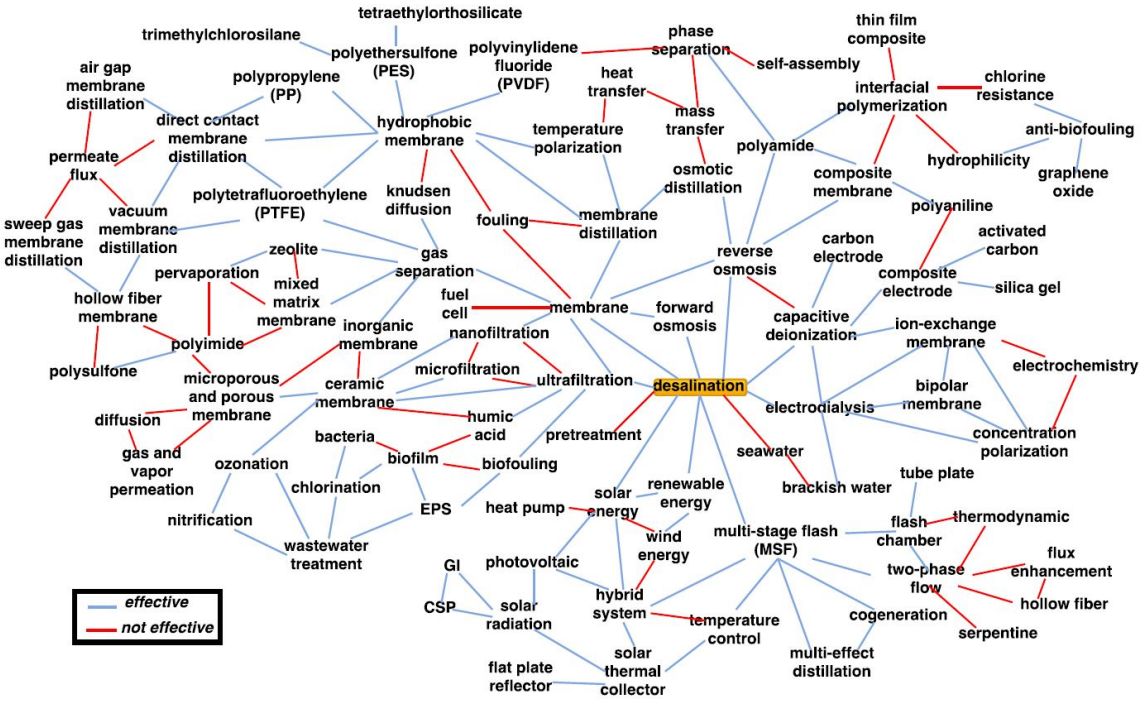

Figure 8. Example network generated by the software

In contrast, WordNet was used to make exploration on "desalination". Even though many relevant terms were retrieved such as chemical process, natural process and physical process, however, few of these concepts are helpful for our practical design activities. The search results using ConceptNet only showed the translations of "desalination" to other languages, e.g. "десалинизација" in Serbian. Furthermore, regarding traditional ideation methods, some research students, in a parallel study, who did not use this software and semantic nets spent a week conducting brainstorming and referring to documents, finally came up with only two concept ideas: reduced-pressure distillation, and hybrid energy system coupling solar panel and PV panel. The low output, in terms of quantity of ideas, of their process is mainly due to the lack of professional expertise in relevant engineering fields. Thus, as shown above, the approach proposed in this paper aims at solving this problem by assisting specialized knowledge retrieval in relevant engineering fields for designers. By using the software we are able to construct the knowledge network shown in Figure 8 within only half an hour which potentially provides at least five feasible engineering solutions with corresponding information about mechanisms, structures and even materials.

6. Conclusion

The approach proposed in this paper aims at assisting design knowledge retrieval for engineering queries by using computer-based data mining method. Current information acquisition methods in engineering design mainly involve interacting with colleagues and referring to documents. Some computer-based knowledge networks such as WordNet and ConceptNet are built either based on hypernym and hyponym or focusing on commonsense. This paper uses the keywords of journal papers as the source for data minning to construct a professional knowledge network. Web crawler, graph algorithm and database techniques are implemented to build the network, and software with a GUI has been developed for user interaction. An engineering design project was used to evaluate this approach. The result indicates that the software is able to retrieve relevant knowledge concepts to support the design task, and can effectively help problem-solving and ideas generation in this case study.

Acknowledgement

The authors would like to acknowledge the Elsevier API for its support in text mining and sample data extraction.

References

Ahmed, S., "An approach to assist designers with their queries and designs", ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, 2006, pp. 359-366.

Ahmed, S., Wallace, K., "Indexing design knowledge based upon descriptions of design processes", DS 31: Proceedings of ICED 03, the 14th International Conference on Engineering Design, Stockholm, 2003.

Altshuller, G., Shulyak, L., Rodman, S., "40 Principles: TRIZ keys to innovation", Technical Innovation Center, Inc., 1997.

Aurisicchio, M., "Characterising information acquisition in engineering design", University of Cambridge, 2005.

Budanitsky, A., Hirst, G., "Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures", Workshop on WordNet and Other Lexical Resources, 2001.

Chakrabarti, A., Bligh, T., "An approach to functional synthesis of mechanical design concepts: theory, applications, and emerging research issues", Artificial Intelligence for Engineering, Design, Analysis and Manufacturing, Vol.10, No.4, 1996, pp. 313-331.

Childs, P. R. N., "Mechanical Design Engineering Handbook", Butterworth-Heinemann Oxford, 2014.

Dev.elsevier.com, Elsevier Developer Portal, Available at: <http://dev.elsevier.com/text_mining.html>, 2015, [Accessed 16.10.2015].

Hastie, T., Tibshirani, R., Friedman, J., "The Element of Statistical Learning: data mining, inference and prediction", Springer series in statistics Springer Berlin, 2009.

Kantardzic, M., "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley & Sons, 2011.

Liu, H., Singh, P., "ConceptNet—a practical commonsense reasoning tool-kit", BT Technology Journal, Vol.22, No.4, 2004, pp. 211-226.

Pedersen, T., Patwardhan, S., Michelizzi, J., "WordNet:: Similarity: measuring the relatedness of concepts", Demonstration papers at hlt-naacl 2004, Association for Computational Linguistics, 2004, pp. 38-41.

Rastegarpanah, A., Mortaheb, H., "Surface Treatment of Polyethersulfone Membranes for Applying in Desalination by Direct Contact Membrane Distillation", Desalination, Vol.377, 2016, pp. 99-107.

Salton, G., Allan, J., Singhal, A., "Automatic text decomposition and structuring", Information Processing & Management, Vol.32, No.2, 1996, pp. 127-138.

Salton, G., Buckley, C., "Automatic text structuring and retrieval-experiments in automatic encyclopedia searching", Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, Chicago, Illinois, USA, 1991, pp. 21-30.

Salton, G., Singhal, A., Mitra, M., Buckley, C., "Automatic text structuring and summarization", Information Processing & Management, Vol.33, No.2, 1997, pp. 193-207.

Seco, N., Veale, T., Hayes, J., "An intrinsic information content metric for semantic similarity in WordNet", ECAI, 2004.

Speer, R., Havasi, C., "Conceptnet 5: A Large Semantic Network for Relational Knowledge", The People's Web Meets Nlp, Springer, 2013, pp. 161-176.

Speer, R., Havasi, C., "Representing General Relational Knowledge in Conceptnet 5", LREC, 2012, pp. 3679-3686.

Suárez, F., Ruskowitz, J., Tyler, S., Childress, A., "Renewable Water: Direct Contact Membrane Distillation Coupled with Solar Ponds", Applied Energy, Vol.158, 2015, pp. 532-539.

Tran, T., Cimiano, P., Rudolph, S., Studer, R., "Ontology-Based Interpretation of Keywords for Semantic Search", The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, Korea, Springer Berlin Heidelberg, 2007, pp. 523-536

Witten, I., Frank, E., "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, 2005.

Zaki, M., Meira Jr., W., "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2014.

Feng Shi, PhD
Imperial College London, Design Engineering
10 Princes Gardens, SW7 1NA London, United Kingdom
Email: f.shi14@imperial.ac.uk