INTERNATIONAL CONFERENCE ON ENGINEERING DESIGN, ICED11
15 - 18 AUGUST 2011, TECHNICAL UNIVERSITY OF DENMARK

A SCALABLE APPROACH FOR THE INTEGRATION OF LARGE KNOWLEDGE REPOSITORIES IN THE BIOLOGICALLY-INSPIRED DESIGN PROCESS

D. Vandevenne [a], P.-A. Verhaegen [a], S. Dewulf [b], J.R. Duflou [a]
[a] Centre for Industrial Management, Department of Mechanical Engineering, K.U.Leuven, Belgium
[b] CREAX N.V., Ieper, Belgium

ABSTRACT

Many applications of Biologically-Inspired Design (BID) are well-known and research is increasingly focusing on methodologies towards systematic BID. However, currently no ideation tool exists that is able to leverage the large textual biological resources in a scalable way to propose a selection of biological strategies that are interesting for a specific design problem under focus. This paper first identifies the main bottleneck preventing the realization of such a scalable BID ideation tool by analyzing the state-of-the-art in systematic BID. It is observed that most work focuses on developing detailed models of, both biological and engineering systems, which enable support during knowledge transfer between the two domains. However, the automated instantiation of these models for a large collection of biological strategies currently remains an open question and domain experts are necessary to complete this time-consuming and expensive task. Therefore, a new approach is proposed that uses a conceptual representation of the biological domain to identify candidate biological strategies as input for the transfer phase.

Keywords: Biological-Inspired Design, Product Aspect, Organism Aspect, Biomimicry, Ideation

1 INTRODUCTION

Biologically-Inspired Design (BID), also known as Biomimicry, Biomimetics, Bio-Inspiration and Bionics, is the discipline where inspiration is taken from nature to solve problems humans encounter. BID is receiving increasingly more attention from research and industry because of the two main advantages the field is often associated with: sustainability and proven performance [1,2]. Supported by the properties of numerous biomimetic examples, of which the publicly available database of Ask Nature [3] forms a comprehensive resource, it is generally accepted that biomimetic products often imply an improvement in sustainability compared to the way classical engineering would tackle similar design problems. Some well-known examples are: saving energy consumption of the Japanese bullet train by inspiration from the shape of the kingfisher’s nose, artificial photosynthesis as a renewable energy source, biodegradable polymers to avoid uncontrollable waste accumulation effects, etc. Furthermore, it is often assumed that the resulting biological strategies are proven in their performance since they are directed by the process of evolution. Being able to directly learn from these biological solutions enables the designer to jump to an already workable solution. Other noteworthy advantages of biomimetic products are their enhanced marketability caused by the entailed green image of such products, by the savings they often imply (e.g. less energy consumption) or by the association to the organism itself (e.g. to swim as fast as a shark with a biomimetic swimsuit); and the higher probability of identifying leapfrog innovations from drawing inspiration from a largely unused biological knowledge domain.

These high expectations, together with the existence of large amounts of biological knowledge, are currently not met with adequate methods and algorithms to enable designers to identify candidate descriptions of biological strategies for biomimetic design from large knowledge repositories. Most existing biomimetic ideas currently originate from spontaneous inspiration. The invention of Velcro is one of the most well-known examples: The inventor, George de Mestral, spontaneously observed the ability of the burrs to attach itself to the fur of his dog. This inspired him to study the phenomenon in detail and to develop the well-known innovation. Another way to integrate bio-inspiration into the innovation process is the employment of a multidisciplinary design team, an approach which, although expensive provides no guarantee for success. A knowledge-based, systematic BID process should generate more biomimetic ideas in less time by reducing the element of chance in ideation. Because of the inherent advantages of biomimetic products discussed above, such a breakthrough could have a positive impact on the overall sustainability of our way of life.

The main intentions of this paper are to argue for the development of scalable methods and algorithms that support BID by leveraging the large amounts of biological knowledge that are currently left undisturbed, to expose the gap in related research in order to achieve such a goal and to outline a new and scalable approach positioned with respect to the current state-of-the-art.

2 THE BID PROCESS SUPPORTED BY A LARGE BIOLOGICAL KNOWLEDGE BASE

In this paper the authors use the term *biological strategy* to refer to the natural phenomenon that can serve as design inspiration. Depending on the specific research approach, different terminology is adopted. For example, [4] distinguish physiology, morphology, behavior and strategy as different viewpoints for biological systems. They define strategy as generic behavior that is exhibited among multiple biological ranks to achieve different goals. In this paper the authors use the term *biological strategy* more generically to describe any solution nature has devised to solve a problem and the term *biological system* for the organism (or parts of the organism) and the relevant elements in its environment to explain the biological strategy.

Currently only about 1.7 million species are named, while the total number is expected to be between 5 and 30 million [5]. Furthermore, from these 1.7 million identified organisms only a small fraction is studied in detail. Nevertheless, there exist many sources, such as books, journals and online resources, where biological knowledge is documented and considering the large work that lays ahead for biologists to completely describe and comprehend all of nature’s phenomena, these sources are expected to keep on growing. However, an important barrier for designers to find relevant biological solutions in large repositories is caused by the very different terminology that engineers and biologists use to describe the systems in their domains. An example, given by [6] explains how a designer, looking to develop a product that is required to be *self-cleaning*, should actually search for strategies where the organism *defends* itself. In this way, for example, the self-cleaning surface of the lotus flower can be identified.

From a literature analysis covering different approaches to the BID process [7], it is found to consist of four main steps: formulating search objectives, searching for biological analogues, analyzing biological analogues and transfer. These four steps are applied in this paper to the systematic BID process that draws inspiration from large textual biological repositories.

- **Step 1: Formulate search objectives**. The specific design problem at hand needs to be captured in a form that allows search in the next step.
- **Step 2: Search**. Searching large repositories of biological strategies requires a scalable, automated approach that allows the identification of a number of candidate biological strategies to be considered in the next step. In order to ensure scalability, requiring manual or expert intervention for integrating each biological document in a structured database is not an option. Furthermore, requiring manual interaction for the evaluation of each biological document in the database during a search also does not scale.
- **Step 3: Filter and Analyze**. An automated search method unleashed on large repositories is expected to generate more than a few candidate descriptions of biological strategies. Therefore, methods and algorithms are required to guide the designer in selecting one or more candidates, e.g. by grouping similar strategies or by ranking to relevance criteria. The retained biological systems should be analyzed in detail to enable the designer to transfer the biological principles in the next step.
- **Step 4: Transfer**. In order to transfer knowledge from the biological source domain to the technological target domain in a systematic way, methodologies or tools can assist the designer in realizing the cross-domain analogy and to come up with a feasible, biologically-inspired technical concept.

3 RELATED RESEARCH

Although the field of biologically-inspired design is still in an early development stage, there definitely has been an increase in research interest over the last decade. This has resulted in new biomimetic applications and in new research approaches towards systematic BID. In this section, existing contributions to facilitate bio-inspiration are positioned in these BID process steps, applied to the premise of leveraging large biological repositories, as discussed in the previous section. It is important to note that the absence of one or more BID steps in a specific research approach should not necessarily be considered as a methodological flaw as it can be the aim to, for instance, only support up to search (step 2) and to leave filtering and analysis (step 3) and transfer (step 4) out of scope. Other reasons for the absence of a view or solution on specific steps in the BID process might be that research is still ongoing or that certain steps of the BID process, e.g. scalable search, are not considered by the authors.

3.1 Related research positioned in the scalable BID process

Ask Nature. The approach of Ask Nature [8] places biomimetic products and biological strategies in a functional, hierarchical taxonomy called the biomimicry taxonomy. The designer looking for bio-inspiration needs to formulate his or her design problem (step 1) in this biomimicry taxonomy. For example, when looking for bio-inspiration for a robot that climbs walls, the problem could be formulated as *move or stay put* at the highest level in the biomimicry taxonomy, as *move* at the second level and as *move on solids* at the third level. Search (step 2) occurs by the indexes manually created according to the biomimicry taxonomy. No aid is provided for analyzing, ranking or filtering the results (steps 3) nor for transfer (step 4). Although this tool does not support these last two BID process steps, the Ask Nature tool currently is one of the most useful aids for designers looking for bio-inspiration. However, the approach relies on manually created entries for each biological strategy, a task performed by Ask Nature and partly outsourced to the biomimicry community on a voluntary basis. In 2009, the database grew on average with about 1 strategy every 2 days and currently holds 1369 strategies in total. Because of the large amount of documented biological knowledge, this manual approach is not expected to scale and the vast majority of biological knowledge is expected to never be manually integrated in this database.

Bridge verbs. Another contribution proposed by [9] attempts to bridge the terminology gap between the engineering and biological domain by means of a systematic, semi-automatic search method that requires the design problem to be expressed in functional keywords (step 1) and then generates biological meaningful bridge verbs and text passages containing them [6,9]. To the best of our knowledge, this is the first and only attempt to algorithmically search for biological strategies in natural language format as a response to a design problem, while recognizing the nuisance of the inherent terminology gap. However, the method has only been documented to be tested on a single resource: the introductory biological textbook ‘Life, the science of Biology’ [5]. Because of the iterative nature of the method, requiring manual evaluation of numerous retrieved pieces of text for each document, the approach cannot be expected to scale to large biological corpora. As stated in [10]: “Even with a single text used as source, there can be an unmanageable number of matches”. Without this manual filtering step, the method is expected to converge to the most used biological terms. In the example given, even with manual filtering, *convert* and *surround* are related as first and second suggestion to the query term *encapsulate* and these same words are the second and fifth proposal for the query term *remove*. The above method was used by [11,12] to generate bio-inspiration and investigated how the retained textual descriptions are used in the transfer phase (step 4). Several types of fixation were found: fixation on certain words, fixation on certain phrases and fixation on certain solutions regardless of the descriptions used as stimuli. Other conclusions are that more support during transfer is required and that ambiguity in the biological description can also be useful during ideation as opposed to detailed information for precise comprehension of the phenomenon. In summary, these studies [11,12] point to the necessity to carefully consider the representation of identified candidate biological strategies in natural language format as input for filtering and analysis (step 3) and transfer (step 4).

BioTRIZ. Another approach proposed by [13] consists of the integration of biological knowledge in the TRIZ methodology [14]. TRIZ, Theory of Inventive Problem Solving, is a collection of tools and techniques based on contradictions to represent problems. In such an approach (engineering) problems as well as (biological) solutions are abstracted to the contradiction in design parameters that characterizes the systems in both domains and the inventive principles that apply for the resolution of the conflicts are identified for reuse later. It is important to note that the identification of these inventive principles for classical TRIZ required an enormous effort of manually analyzing a huge number of patents, over three million, and these resulting inventive principles are not automatically updated for recent inventions. Such an approach was pioneered by [13] to include bio-inspiration in the TRIZ methodology by developing the BioTRIZ matrix. This matrix has six parameters based on the mantra: things (substance, structure) do things (requiring energy and information) somewhere (in space, time). By analyzing the biological resolution of 2500 conflicts, which is less than a fraction of the data behind classical TRIZ, it was found that nature is also bound by the 40 inventive principles from classical TRIZ [13]. However for a specific contradiction different inventive principles are used in biology. Therefore, the authors of [13] repositioned them into a new matrix, named BioTRIZ. This approach is positioned in the premise of leveraging large biological repositories in the BID process as follows. In step 1, the problem needs to be formulated into a classical TRIZ contradiction, which is then reformulated into a BioTRIZ contradiction between the six parameters discussed above. This BioTRIZ contradiction then leads the designer to the inventive principles learned from the manual analysis of 2500 contradictions in 500 biological phenomena. To enable a search in large biological repositories, in step 2, the involved contradictions and inventive principles need to be identified beforehand for each biological description in the corpus, a task which does not scale well. The BioTRIZ approach is designed to answer with abstract inventive principles from biology and therefore, by itself, does not provide support for transfer, in step 4. Furthermore, using such a (Bio)TRIZ-based approach requires extensive training and domain knowledge from both the engineering and biological domain.

Functional basis. The approach proposed by the authors of [4] represents both the engineering and biological system in a functional model, using the functional basis, in order to facilitate concept generation. First, in step 1, the customer needs are expressed in a functional model. In step 2, a suitable biological system is identified by automated methods based on similarities in the functional model. This requires a database containing biological functional models to be available. Although a detailed methodology to guide the biological modeling process is provided, it is explained that functional modeling of a biological system is a time-consuming, iterative process requiring both engineering and biological expertise. This complexity of the biological functional modeling process makes automation towards large biological repositories unlikely. However once a low number of viable candidate strategies are identified and analyzed in step 3, this approach can facilitate technical bio-inspired concept generation during transfer by instantiating a detailed, similarly structured functional model for the systems in both domains. Therefore, the rationale behind the research is strongly focused towards step 4, transfer, and the bottleneck is situated in step 2, searching large databases.

DANE (Design by Analogy to Nature Engine) is an interactive, knowledge-based design environment that provides access to a small design case library containing Structure-Behavior-Function (SBF) models of biological and engineering systems [15]. 22 complete models of biological systems are documented. The SBF models are used for capturing the functioning of biological systems, where functions serve as indices to organize the knowledge about behavior and structure. DANE uses rich content (text, schemes, images, diagrams, etc.) to represent the models to the designer. In step 1, the design problem needs to be represented by instantiating at least parts of such an SBF model. In order to help the designer to find biological systems, in step 2, these biological systems also need to be represented in SBF models and stored in the design case library. From tests with interdisciplinary teams of students with biological, mechanical engineering and various other backgrounds, this modeling task has been proven to be difficult to perform manually and to be time consuming (40 to 100 hours per model). Automating this task to leverage large textual biological corpora is infeasible in the near future. However, by forcing designers to analyze the candidate biological strategies in detail (step 3) and to construct SBF models for biological and engineering systems (step 4), a detailed understanding of their functioning can be achieved and the obtained abstraction can facilitate knowledge transfer between both domains as well as communication in multi-disciplinary design teams. Furthermore, the ability to connect functions across different biological scales (e.g. relating the muscular contraction function to animal level functions like movement or feeding) is a promising feature for transferring biological knowledge. The main contribution of this approach is again situated in step 4, transfer, and the main bottleneck is again the integration of large amounts of biological knowledge to increase relevant search results.

IDEA-INSPIRE is a software tool based on a database of engineering and biological systems both represented by the SAPPhIRE model of causality for engineering and biological systems, interlinking State, Action, Part, Phenomenon, Input, oRgan, and Effect by causality [16]. This is also a model based on Function, Behavior and Structure which are related as follows. Function is seen as the intended effect of a system and behavior as the link between function and structure at a given system level. The model allows using multiple levels of abstraction to explain how a system works to fulfill its goals. In step 1 the designer looking for bio-inspiration expresses his problem by instantiating parts of the SAPPhIRE model for a direct or indirect search. For a direct search, the intended action can be represented by a verb-noun-adjective triplet. For an indirect search, constructs linked to the desired action, e.g. associated state changes, can be instantiated. In order to search a large set of documented biological strategies or systems, manual instantiation of one or more SAPPhIRE models per biological system would require an enormous effort to populate a database. A database containing 100 entries of models of plants or animals is reported by [16]. Assuming steps 2 and 3 would result in a low number of viable candidate biological strategies, one or more SAPPhIRE models could be instantiated and extensive support for transfer, in step 4, is possible with the SAPPhIRE model, SAPPhIRE guidelines and the rich information (text, images and video) that the IDEA-INSPIRE software is intended to present to the designer. The SAPPhIRE transfer model and guidelines are validated by modeling the functionality of 20 biomimetic examples in 80 SAPPhIRE instance pairs [7].

3.2 Discussion

First, the authors observe that there is only one approach [6] that attempts to integrate biological knowledge in natural language format. Although the intent of the approach is in accordance with the goals set for this paper, the described method is found not to scale for searching large knowledge repositories (step 2). Two other approaches, one based on a hierarchical functional taxonomy [3] and the other based on the TRIZ methodology [13], also focus on assisting biomimetic search in step 2. However they require manual expert contribution for the integration of each biological strategy and therefore do not scale for leveraging large existing knowledge repositories. To the best of our knowledge, no method exists that enables a search for candidate strategies from large biological knowledge repositories. Table 1 gives an overview of the database sizes and content handled by the current state-of-the-art.

Table 1. Overview of existing database sizes and content

Method	**Size and Content**
Ask Nature	1369 detailed descriptions of biological strategies
Bridge verbs	1 biological introductory handbook
BioTRIZ	2500 conflicts, from an analysis of 500 biological phenomena
Functional basis	30 models of biological phenomena
DANE	40 of which 22 complete SBF models of biological systems
IDEA-INSPIRE	20 biomimetic examples (engineering and biological systems) 100 biological strategies about motion in nature

All other approaches focus on step 4 (transfer) by composing detailed models (functional basis, SBF, SAPPhIRE) of the biological and engineering systems to help the designer to make cross-domain analogies. Although one might argue that such models can greatly increase the effectiveness of search in step 2, the instantiation of these models is a labor-intensive task requiring both biological and engineering experts and is therefore not scalable for large knowledge repositories. Therefore, these transfer models should be constructed manually only for the identified candidate biological strategies in step 3. The automated component of step 3, necessary to filter a large number of identified candidate strategies after search, is never discussed in the above overview of related work because it requires the existence of a scalable search method. The manual component of step 3, a detailed analysis of the candidate biological systems, is implicitly present for all methods that focus on transfer, in step 4, in order to instantiate transfer models.

The one feature that all contributions share is the presence of a functional component in their methods and models. Variants are the use of functional keywords for search, searching a functional taxonomy, the functional basis model, the Function-Behavior-Structure model, the SAPPhIRE model to capture the functionality of systems and BioTRIZ where the biological functions are related to TRIZ conflicts and inventive principles. Therefore, the functions of biological and technical systems will be a central component of the proposed approach, outlined in the next section. This approach will focus on step 2, scalable search for biological candidate strategies, because it is the identified bottleneck in the above BID process.

4 PROPOSED APPROACH

The authors propose to compose a conceptual representation of the relevant knowledge about the biological domain. Automated characterization of biological strategies, and of the involved organisms, enables a scalable search over large databases, only limited in size by the amount of biological natural language knowledge that exists. In this way, the proposed approach addresses the bottleneck in the current state-of-the-art that prevents designers to identify relevant biological knowledge from large repositories in natural-language format. The analysis of related work points out that the functions of biological systems should be central in such a conceptual representation. The detailed transfer models also focus on structure and behavior and how these are interlinked to realize a function. The authors argue to postpone such a time-consuming manual analysis to the fourth step, transfer, and to focus on function to identify candidate biological documents as a source of inspiration.

In the presented research, the concepts that represent the biological domain are called Organism Aspects (OAs), analogous to previous research by [17] where a conceptual representation about the engineering domain is composed with concepts named Product Aspects (PAs). These Product Aspects are automatically generated, fine-grained concepts, based on the analysis of term occurrences in patents [17]. A software tool called PAnDA (Product Aspects in Design by Analogy) implements these PAs and is validated to effectively assist in the ideation phase [18] by relating a product under focus to relevant other products. Examples of Product Aspects are: heating, cutting, cleaning, etc. The concepts that represent the biological domain, OAs, are generated from a similar analysis of biological documents. As there are no "biological patents", existing biological knowledge in natural language format is used, as explained in Section 4.1. Our approach focuses on steps 1, 2 and 3a of the scalable BID process, as illustrated by Figure 1 and explained in detail in Section 4.2.

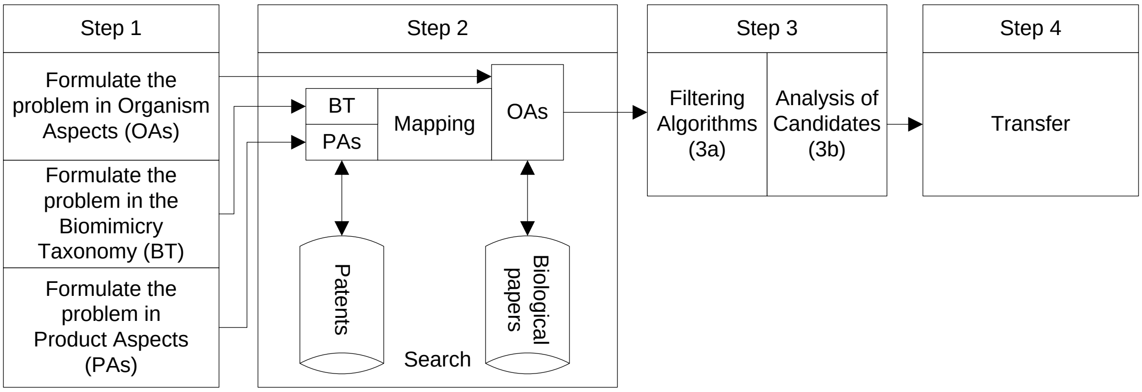

Figure 1. Proposed scalable BID process

4.1 The construction of Product and Organism Aspects

In order to construct PAs and OAs, knowledge repositories are required. These repositories should contain descriptions that capture relevant information about products for technical systems and about organisms for biological systems. For technical systems, patents are randomly sampled from the EPO Worldwide Patent Statistical Database (PATSTAT) database [19]. Patents are relevant sources because they each describe a new idea and have been critically reviewed by experts. The main disadvantage of patents is the common practice of intentional obfuscation or generalization of the content for reasons related to economical competition. For biological systems, scientific descriptions of the natural phenomena are most suited. Such documents do not suffer from the intentional ambiguity, like the patents do. Furthermore, scientific descriptions are peer reviewed, structured and focused on a single biological topic. However, most large repositories are proprietary and publishers are typically adverse towards mass-processing of the data on which they possess copyrights. For the construction of the conceptual representation, a representative sample of the domain is sufficient, because it has been observed that, once a critical number of documents is analyzed, this conceptual representation remains stable. The authors are planning to investigate the suitability of open-access sources such as PubMed Central [20] and to determine the feasibility of applying webcrawling to methodologically search the web for suitable sources. When using open-access repositories, filtering algorithms will be essential to guard the relevance of the retained biological documents. For example, documents about cancer research on rats and about the mapping of the genome of mice will be omitted, since they are not relevant for BID. When using webcrawling, relevance feedback during the crawling process is equally important. Filtering and relevance feedback algorithms could be based on: the size of the documents, the presence of a focus organism in the descriptions, the keywords of a paper, etc. When biological books, like [5], are considered as sources, these need to be split into their coherent parts without manual intervention. Several automated approaches for text segmentation exist, e.g. [21,22], which can be applied for splitting biological books.

To construct OAs, a similar procedure can be adopted as for the construction of PAs, which is detailed in [17] and illustrated by Figure 2. The core of the procedure involves the representation of a corpus, a collection of biological documents, according to the Vector Space Model (VSM) in a Document-Term Matrix (DTM) which is then subjected to Principal Component Analysis (PCA) [23]. In such a DTM each coordinate represents the number of times a term occurs in a document. PCA allows extracting a given number of Principle Components (PCs), of which the first PC is the linear dimension oriented in such a way that it explains the maximum amount of the variance in the data set. Each succeeding PC represents as much of the remaining variability as possible, taken into account that all PCs are orthogonal to each other. After PCA, all terms are expressed in a smaller number of uncorrelated variables or PCs, resulting in a term-PC matrix. Furthermore, it has been demonstrated in [17] that Varimax rotation [24] facilitates the interpretability of the resulting PCs. After rotation, the PCs are called Organism Aspects and automatically represented by a number of ranked terms. These groups of ranked terms are manually interpreted by different domain experts to label the OAs.

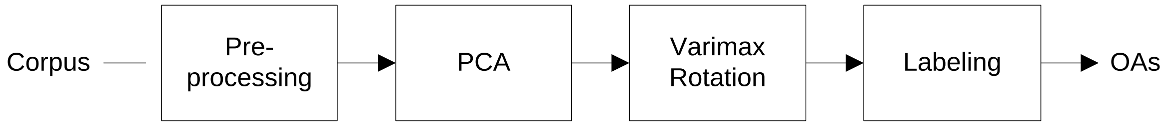

Figure 2. The construction of OAs

Without preprocessing of the corpus, applying the above procedure on biological documents would result in many concepts that are useless for the envisaged application. For example, the concept roman numerals represented by the following terms: I, II, III, IV, V, etc. will be identified because these terms are highly correlated and therefore form a concept. Furthermore, because of the high dimensionality of the DTM, dimension reduction with PCA would also be a computationally very expensive task. To bring out the interesting structure in the corpus, systematic term filtering is applied. For example, terms grammatically used as verbs are retained by Part-Of-Speech (POS) tagging [25] because they are likely to be related to the function of the system, while other terms, such as the names of organisms, are removed in order to eliminate taxonomic structures from the resulting OAs. Another preprocessing step involves a manual filter of the above retained terms by relevance for knowledge transfer. In this way, for example, the verb to become is eliminated and the verb to protect is retained. Finally, also the typical preprocessing steps related to information retrieval are required to guarantee accuracy of the results, computational feasibility of PCA and to alleviate the manual filtering and labeling discussed above. Such preprocessing steps are term weighting, e.g. Term Frequency Inverse Document Frequency (tf-idf) [26], normalization to account for the differences in document size, and stemming [27] to further reduce the dimensions of the DTM.

4.2 Organism Aspects enabling a scalable BID process

For step 1, problem formulation, three scenarios are envisaged. A first method consists of a direct selection of Organism Aspects, which are high-level functional concepts that characterize organisms and their biological strategies. A second method to express the design needs is based on the hierarchical functional taxonomy of Ask Nature [3]. Although some biological knowledge is required for this task, informal experiments with mechanical engineering students have given indications that this method is effective in ideation, however currently limited by the size of their database. A third method is based on the Product Aspects that characterize products. The designer could generate these PAs with the PAnDA tool discussed above, or he or she could select them manually.

For step 2, these problem formulations need to be connected to the automatically generated OAs to enable a scalable search. To connect the first type of problem formulation to OAs, no mapping is necessary as OAs are directly used. To connect the second and the third type of problem formulations, mappings are required from the high-level functional OAs to the hierarchical functional taxonomy of Ask Nature, or to the PAs. Currently, the authors envisage two methods for such mappings, an automated method that is based on partly shared vocabulary in the terms related to the extracted concepts and a manual method that is performed by experts. Manual mapping does not impede with overall scalability, because it is performed on a manageable number of high-level concepts, e.g. a couple of hundred which is an analogue number to the number of concepts used in [17] to represent products in the technical domain. It is envisaged to combine both the automated and manual methods. The result of step 2 is a set of selected biological documents to facilitate the next two steps.

Consider, for example, the design problem of developing a robot which is able to move on walls and ceilings. In the first scenario described above, the OAs are directly selected by the designer. Intuitive functional OAs for this problem are “move” and “attach”, two concepts that are likely to be identified by the process depicted in Figure 2. An OA like “attach” encompasses terms like: attach, join, fix, bind, glue, suck, etc. The strategy specific terms (e.g. bind, suck, and glue) are grouped by their co-occurrence with more common words (e.g. attach and join). In the second scenario, the design problem is formulated in the hierarchical functional taxonomy of Ask Nature [8]. At the highest level “move or stay put” is the appropriate category to select, at the second level “attach” and at the third level “temporarily”. It is likely that OAs are identified which can be linked to such categories of the biomimicry taxonomy, as described above. In the third scenario, such OAs are connected to relevant PAs. As there are many products that “move” and “attach”, such concepts are also expected to emerge among the PAs, however, they group engineering related terms. A PA like “attach” can group specific engineering terms like glue, weld, solder, crew, clamp, etc. by their co-occurrence with general terms like attach and join. An advantage of using PAs in problem formulation is that these can be generated automatically for a product under focus. For all three scenarios, once the relevant OAs are identified, these will lead to the biological strategies from which they are generated in step 2, search.

In order to deal with a potentially large number of candidate strategies generated in step 2, additional algorithms organize the results in ways that facilitate semi-automated filtering, in step 3. A required supporting algorithm identifies the focus organism(s), typically one, per description of a biological strategy. However, two or more focus organisms are possible, for instance in a biological function achieved through a symbiotic relationship. The identification of focus organisms is a feasible task: related research in bio-informatics, e.g. [28], pursues the identification of focus organisms in biomedical literature in order to disambiguate other named entities, like proteins. Focus organism identification enables the grouping of similar biological strategies for taxonomically closely related organisms, see Figure 3. This can group, for example, different biological documents explaining how various species of birds fly. Next, ranking can be accomplished by considering the biodiversity of the retrieved and grouped biological strategies. This tests the assumption that strategies which occur for more, and for more taxonomically diverse, organisms entail more evidence for proven design by natural selection and therefore are more interesting to consider. For example, both the lotus plant and the butterfly, taxonomically very distant organisms, protect themselves from accumulating dirt on their leaves or wings by a hydrophobic self-cleaning surface. This enables the ranking of the results to predict their usefulness in solving problems before the transfer phase, a functionality that currently does not exist based on an objective parameter like biodiversity. Semi-automated filtering, step 3a in Figure 1, results in a number of ranked and grouped biological strategies documented with scientific papers. This forms the input for step 3b, manual analysis of the remaining candidate strategies, a necessary preparation for transfer.

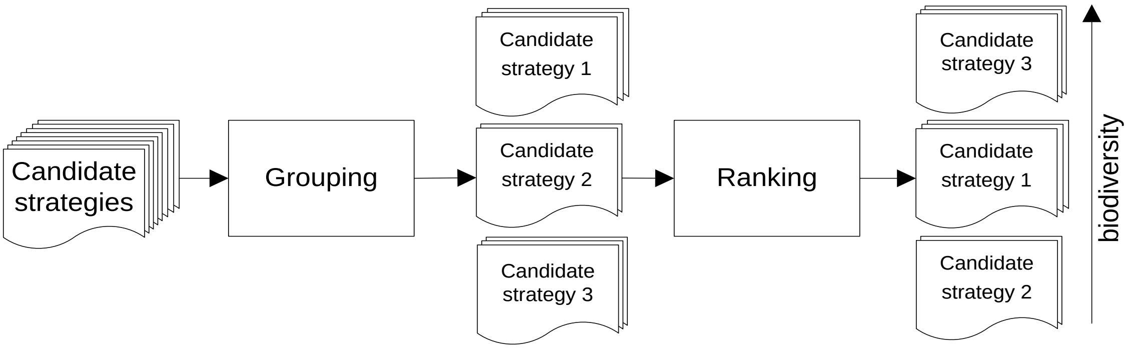

Figure 3. Filtering candidate biological strategies

In step 4, detailed transfer between both knowledge domains takes place in order to result in feasible technical concepts. Any of the above methods that argue for the construction of detailed models could be applied in order to support the designer in making a detailed design by analogy. Because of the manageable number of viable biological candidate strategies, it is now feasible to manually construct such models for relevant biological strategies that originate from a large knowledge repository.

5 CONCLUSION

This paper argues for the development of scalable methods and algorithms that support BID by leveraging the large amounts of biological knowledge in natural language format. From an analysis of related work, it is observed that detailed transfer models are too complex for automated mass instantiation and that there is no existing contribution that effectively identifies candidate strategies for transfer from large databases. Therefore, a new scalable system is proposed, based on a conceptual representation of the biological domain, that focuses on the search and filtering of large number of biological documents in the context of a given design problem. The resulting strategies should be considered as input for the transfer phase, where feasible technical concepts are generated.

REFERENCES

[1] Benyus J. M. *Biomimicry: Innovation Inspired by Nature*, 1997 (Harper Perennial). ISBN: 0060533226

[2] Bar-Cohen Y. *Biomimetics: biologically inspired Technologies*, 2006 (CRC/Taylor & Francis). ISBN: 0849331633

[3] Browse the biomimicry taxonomy, Ask Nature, The Biomimicry Institute, http://www.asknature.org/browse, last accessed on 10/01/2011

[4] Nagel J.K.S., Nagel B.I., Stone R.B. and McAdams D.A. Function-based, biologically inspired concept generation. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 2010, 24(4), pp521-535

[5] Purves W.K., Sadava D., Orians G.H. and Heller H.C. *Life:The Science of Biology*, 2001 (W.H.Freeman & Co Ltd) ISBN: 0716726297

[6] Chiu I. and Shu L.H. Biomimetic design through natural language analysis to facilitate cross-domain information retrieval. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 2007, 21(1), pp45–59

[7] Santori J., Pal U. and Chakrabarti A. A methodology for supporting "transfer" in biomimetic design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 2010, 24(4), pp483-505

[8] Ask Nature, the Biomimicry Design Portal, available from http://www.asknature.org, last accessed on 10/01/2011

[9] Chiu, I., and Shu, L.H.. Bridging cross-domain terminology for biomimetic design. Proc. ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference., Paper No. DETC2005/DTM-84908, Long Beach, CA, September 2005, pp24–28

[10] Shu L.H. A natural-language approach to biomimetic design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 2010, 24(4), pp507-519

[11] Mak T.W. and Shu L.H. Use of biological phenomena in design by analogy. *Proceedings of*

ASME Design Engineering Technical Conferences, DETC2004/DTM-57303, Salt Lake City, Utah, September 2004

[12] Mak T. and Shu L. Using descriptions of biological phenomena for idea generation. *Research in Engineering Design*, 2008, 19(1), pp21–28

[13] Vincent J.F.V., Bogatyreva O.A., Bogatyrev N.R., Bowyer A. and Pahl, A.K. Biomimetics: its practice and theory. *Journal of the Royal Society: Interface*, 2006, 3(9), pp471–482

[14] Altshuller G.S., *Creativity as an Exact Science: The Theory of the Solution of Inventive Problems*, 1984 (Gordon and Breach Science Publishers) ISBN: 0677212305

[15] Vattam S., Wiltgen B., Helms M., Goel A. and Yen J. DANE: fostering creativity in and through biologically inspired design. *First International Conference on Design Creativity*, Kobe, Japan, November 2010

[16] Chakrabarti A., Sarkar P., Leelavathamma B. and Nataraju B.S. A functional representation for aiding biomimetic and artificial inspiration of new ideas. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 2005, 19(2), pp113-132

[17] Verhaegen P., D'hondt J., Vandevenne D., Dewulf S. and Duflou, J. Identifying Candidates for Design-by-Analogy. *Computers in Industry*, 2011

[18] Verhaegen P., Peeters J., Vandevenne D., Dewulf S. and Duflou J. Effectiveness of the PAnDA ideation tool. *Proceedings of the TRIZ Future Conference*, November 2010, pp59-66

[19] EPO Worldwide Patent Statistical Database (PATSTAT), available from http://wiki.epfl.ch/patstat, last accessed on 10/01/2011

[20] PubMed Central Homepage. http://www.ncbi.nlm.nih.gov/pmc/, last accessed on 10/01/2011

[21] Reynar J.C. *Topic Segmentation: Algorithms and Applications*, Ph. D. thesis, Univ. of Pennsylvania, 1998

[22] Ji X. and Zha H. Domain-independent Text Segmentation Using Anisotropic Diffusion and Dynamic Programming, *Proceedings of Special Interest Group on Information Retrieval* (ACM SIGIR), 2003, pp322-329

[23] Skillicorn D. *Understanding Complex Datasets: Data Mining with Matrix Decompositions*, 2007 (Chapman & Hall/CRC). ISBN: 9781584888321

[24] Kaiser H.F. The Varimax criterion for analytic rotation in factor analysis. *Psy- chometrika*, 1958, 23(3), pp187–200

[25] Charniak E. "Statistical Techniques for Natural Language Parsing". *AI Magazine*, 1997, 18(4), pp33–44

[26] Salton G. and Buckley C. Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 1988, 24 (5), pp513-523

[27] Porter M.F. An algorithm for suffix stripping, *Program*, 1980, 14(3), 130-137

[28] Kappeler T., Kaljurand K. and Rinaldi F. TX Task: Automatic Detection of Focus Organisms in Biomedical Publications. *BioNLP workshop*, NAACL/HLT, Boulder, Colorado 2009