Context of Text: Concepts for Recognizing Context of Acquired Knowledge from Documents

N. Madhusudanan, B. Gurumoorthy, Amaresh Chakrabarti

▶ **To cite this version:**

N. Madhusudanan, B. Gurumoorthy, Amaresh Chakrabarti. Context of Text: Concepts for Recognizing Context of Acquired Knowledge from Documents. 14th IFIP International Conference on Product Lifecycle Management (PLM), Jul 2017, Seville, Spain. pp.632-641, 10.1007/978-3-319-72905-3_56 . hal-01764185

HAL Id: hal-01764185
https://inria.hal.science/hal-01764185

Submitted on 11 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Context of text: concepts for recognizing context of acquired knowledge from documents

Madhusudanan N, B Gurumoorthy, Amaresh Chakrabarti

Technological Data Analytics Laboratory for Smart Network Enabled Manufacturing,
Centre for Product Design and Manufacturing, Indian Institute of Science, Bengaluru, India.
{madhu, bgm, ac123}@cpdm.iisc.ernet.in

Abstract.

Documents contain expert knowledge that can be potentially reused across products and lifecycles. In this research, the aim is to reuse diagnostic knowledge from the assembly phase in the next design/planning stage, by acquiring knowledge from documents. While the acquisition itself is a separate research problem, an important part of the acquired knowledge is the context in which it is expressed in the documents. This context dictates the set of situations to which the knowledge applies. In this paper we study various methods from literature that address this challenge in different domains. We highlight current challenges faced in this work. Two possible means to identify sources of context that are built upon previous work are then discussed. Based on one of them, a method for understanding the context of issues in documents is then proposed. The implementation and evaluation of the method are ongoing.

Keywords: Context, knowledge acquisition, assembly situation, assemblability factors

1 Introduction

Knowledge is a core resource of an organization. Its capture, storage and reuse are perceived to be useful in the current day product development [1]. Product Lifecycle Management (PLM) systems do enable accessibility of information across the entire product lifecycle. However, there is also a need to enable PLM systems for connecting pieces of knowledge across the lifecycle [2]. In particular, extracting knowledge created during one product lifecyle to be used in the lifecycle of subsequent products is of considerable interest and an open problem. This paper is part of on-going research on reuse of diagnostic knowledge of assembly and manufacturing issues across the product lifecycle. The objective is to acquire knowledge about problems in assembly, from documents that are typically generated during the lifecycle of a product, that contain the knowledge. This knowledge is intended to be reused during subsequent assembly design/planning stages. It is hoped that such a reuse will prevent repetition of similar issues, leading to products that are less difficult to assemble. The proposed method of knowledge acquisition is shown in Fig 1; it can be noted that there is a gap between the acquired knowledge and the knowledge to be applied. The knowledge is acquired in the

context of the document, but is applied in the context of a current assembly plan. These two processes – acquisition and application – have to be matched in terms of the right context; this is elaborated further in the following sections.

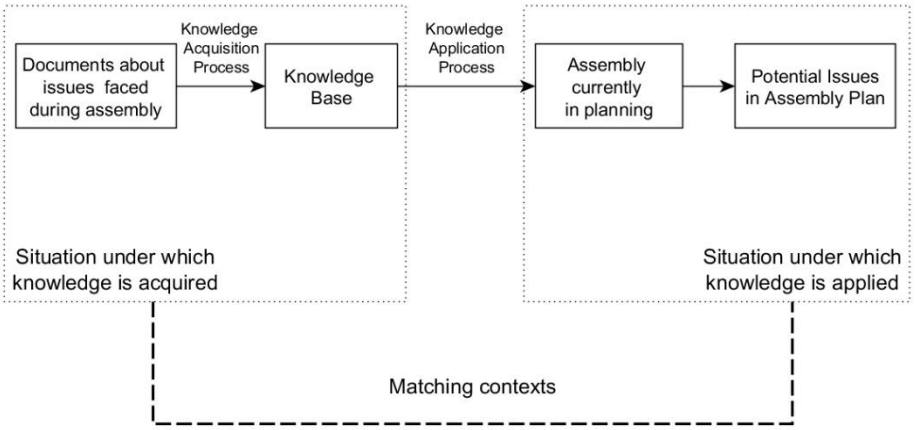

Fig. 1. The two sides of knowledge – acquisition and application; the need to match contexts arises from the two situations on either sides

2 Context

Many different definitions exist for context. As noted in Brézillon [3], literature struggles to agree on a singular definition, since it varies with usage. For the purposes of this work, we find at least two of these useful for our purpose: '*a set of preferences or beliefs*', and '*assumptions under which a statement is true or false*'. From a computing perspective, Henricksen *et al.* [4] defined it as '*Circumstances or situation in which a computing task takes place*'. In our research, knowledge about issues are assumed to be true only in the presence of a set of assumptions or conditions. The latter definition about circumstances dictates that a given chunk of knowledge is usable in a specific situation only.

For example, consider the following issue description from an assembly related text – "*Another disadvantage, pop rivets, being hollow, also introduced the need for sealing each of the holes in the installed rivet heads*". In this case, the knowledge acquired is there could be a need to seal the holes in installed rivet heads, if one type of rivets are used. This knowledge is useful only in the context of assembly operations where 'rivet' is used, and in particular, the 'pop rivet' is used.

Need to recognize context

In particular, the need of this paper is to ensure that knowledge acquired in one context can be applied in the right context. Brézillon [3] clearly stated that context of use must be elicited for a Knowledge Based system. Hence a channel is needed for ensuring that the context of knowledge during acquisition matches the context while using it, as shown in Figure 1.

3 Literature

Literature discusses various aspects of context in text, as well as methods and tools to represent context.

Pomerol and Brézillon [5] discuss and identify relationships among context, knowledge and contextualized knowledge. A theory that retains the context structure while mapping natural language sentences to logical form has also been developed by Kamp (mentioned in [5]). It is understood that the lack of explicit representation of context is a factor leading to the failure of Knowledge-Based Systems (KBSs) [3]. The contextual component of knowledge to be acquired from human experts cannot be missed, and the context of use should be acquired. Held *et al*. [6] stress on the need for context to have a good common representation format.

Henricksen *et al*. [4] say that context is temporally influenced, is imperfect, has various abstraction levels, and is interrelated amongst different context information. They define a classification, and discuss structural constraints of associations that can be modeled for a domain. But these cannot be used for realising context from text. The domain of pervasive computing, from the previous decade, has a large corpus of work with regard to context. For example, Dey and Abowd [7] have developed a Context toolkit. They highlight distinctions between context-aware and traditional desktop applications such as multi-source input, requirement of additional levels of abstraction, and independence of the context system from a single application. The toolkit itself consists of widgets, aggregators, and interpreters following the above distinctions. Bouzeghoub *et al*. [8] combine process-oriented and ontology based context management systems to identify situations from data. Categorical data such as "User, Activity, Environment (computing/physical), Device, Location and Time" are used in the work to recognise context of situations in real time. However, this may be more useful at the situation of the application, rather than being used on text sources.

The field of natural language processing has also addressed context for its own purposes. For example, context-vectors [9] have been used to measure similarity in text. Concepts such as Vector Space Models (VSM) represent text using vectors [10]. The bases for using these representations are hypotheses such as statistical semantics hypothesis, bag-of-words hypothesis, distributional hypothesis, etc. For example, distributional hypothesis predicts that similar words occur in similar contexts. Among many other applications, document retrieval is a popular one. Although VSMs form a basis for representation of context, it is still very low level – at words and pairs of words and is not organized with respect to meaning. It is still unclear as to how to use these representations as context for knowledge. Also, we have text documents only at one end of

the comparison. On the other side, it is not exactly a text document being used, but a mixed set of not-so-detailed text and numerical information. This mismatch makes it all the more challenging to use the above set of methods, without further extensive study. Finally, the context is about issues, whose sizes are, by nature, very small compared to those of the entire text documents.

4 Possible indicators for context

As mentioned in the previous section, the need to define context of knowledge is influenced by its use. The common representation format mentioned by Held *et al* [6] plays an important role in this activity. Hence, for the purposes of defining context for knowledge, we explore two concepts from previous research that may be potentially useful - the assembly situation model and the five important factors in assembly operations. A comparison of these two methods to choose the more suitable method is then presented in Section 5.

4.1 Assembly Situation Model

As shown in Fig. 1, situations are the means for describing knowledge, at the acquisition and application stages. Ye *et al*. (referred in [8]) define situations for a mobile computing environment as "external semantic interpretations of low-level context, permitting a higher-level specification of human behaviour in the scene and the corresponding system services". As part of earlier research by the same research group, an information model for representing assembly situations was developed [11]. It is called the assembly situation model (ASM), and is shown in Fig. 2. In its elementary form, an assembly task is represented as transitioning from an initial state of unassembled parts to a final assembled state of parts, via an assembly process. The dotted box around the first assembly process (Parts A, B; Process 1; Subassembly AB) shows such an elementary ASM. This model for a single task can be extended to the assembly of an entire assembly tree. As shown in Fig 2, Process 2 combines Part C, Part D and Part E, and Process 3 combines Subassembly AB, Subassembly CDE and Part F, resulting in the assembled product ABCDEF.

The model simplifies the representation of information related to an assembly process. Information about all the parts prior and post their integration into the assembly is available (including geometric and non-geometric information from CAD). Information of possible assembly sequences as well as information about the processes (such as pre-processes, process parameters and description of steps) can also be represented.

With the above representation, the ASM ensures that it is possible to attribute a piece of knowledge to a particular step in the assembly. Otherwise, the only representation of an assembly would have been the final assembled product, that is typically represented by CAD models at the design/planning stage.

4.2 Factors affecting assemblability

In order to understand what information is available at the usage stage of knowledge, a universal set of all necessary information is needed. However, there is currently neither a classification nor a master-list of such information. In order to overcome this disorganized status, the work of Santhi *et al*. [12] is used. They propose that there are five factors that affect assemblability, namely part, person, process, tool and environment.

Note that these factors may be interpreted as representing

a) The different parts of an assembly process that interact together, and

b) A classification scheme for information related to the assembly process.

Hence, it is possible to use the above set of factors themselves, in combination with the assembly situation model, to identify the location of a specific piece of knowledge in the assembly process.

5 Proposed method to recognise context

In the previous sections, two primary means for capturing assembly related information have been proposed - namely the assembly situation model and the assemblability factors. The ASM is associated with an existing assembly situation and its purpose is to construct a usable information model of a situation from available CAD and process data. This makes it more useful at the utilization stage of the knowledge when a situation is available at hand to apply the knowledge upon. However, in this paper, we are placed in the knowledge acquisition stage, where the available information is in text format. Due to this, the assemblability factors look more appealing from the perspective of identifying contextual clues from text. They can be considered as the informational aspects of an assembly process.

Also, similar to context vectors, the context of discussion can be based on the words that are present in the neighbouring text. This argument can be extended also to other parts of the document, such as the title of the document, the section heading of the current text, and so on. Based on these points, the following method is proposed to capture the context of the knowledge being acquired (also see Figure 3):

- If the knowledge about issues is present at a specific location indicated by sentence index S_i in the text, then look at its document title and current section heading.
- Also, store the surrounding text of S_i within a window of predefined distance, say, for example, a 10-sentence window on either sides.
- List all the nouns and verbs in the text window, as well as in the document and section titles.
- Classify these words into the five context containers i.e. assemblability factors. For this, the closest of the categories to which each entity from the previous step relates to, must be identified. Hence, words such as '*rivet*', '*glue*' etc. would be closer to process and tool; '*temperature*', '*humidity*' etc. are closer to environment; and '*shoulder*' and '*head*' would be closer to operator.

- The above values for the context containers constitute a signature of context for the issue-related knowledge under consideration.

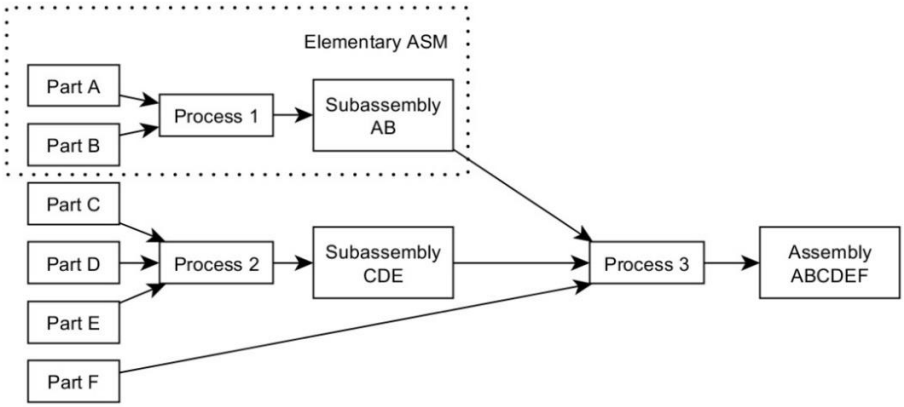

Fig. 2. Example of an assembly situation model for a six part assembly. The dotted portion shows an elementary ASM

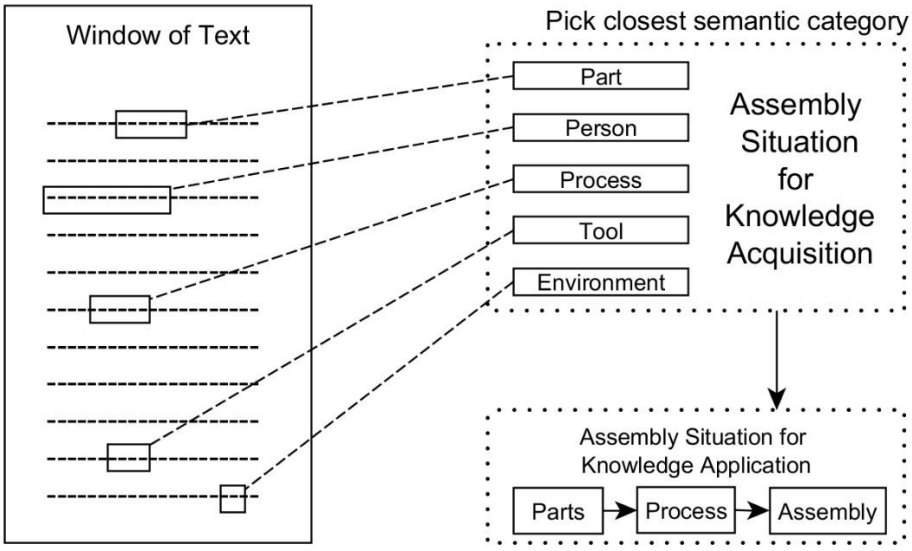

Fig. 3. Proposed method and scheme for identifying context of situation

To illustrate the fourth point above, words intended for each factor are evaluated for similarity with the title of the factors. This is shown in Table 1. WordNet-based '*lin*' similarity measures have been evaluated using the WS4J web-interface (http://ws4jdemo.appspot.com).The closest appropriate synsets (word sense) have been chosen by the tool itself for the highest score, and not by researcher.

It is observed that for the words '*wing*', '*machine*', '*rivet*', and '*sealing*', the decision to categorize them is straightforward. However, with the word '*hand*', the initial category of choice is '*part*'. Due to the word-sense being automatically chosen on the basis of the highest score, the word-sense chosen for '*part*' was '*part#n#3*'('*Portion of natural object*'). When it was evaluated with the more appropriate sense of '*part*' (part#n#2: '*Something less than the whole of a human artifact*'), the similarity with '*hand*' was 0.1561. Similarly, similarity of '*hand*' with '*process*' (process#n#5: '*a natural prolongation/projection from a part of an organism*') was initially 0.4972. But after changing to the more appropriate sense (process#n#1: '*particular course of action intended to achieve a desired result*'), it was 0.3639. These are indicated in parentheses in the table.

Hence, use of the correct word sense is a minor yet influential factor in the classification of words. We now describe a more detailed example of how this method can be used.

Word	**Part** (*'part'*)	**Person** (*'person'*)	**Tool** (*'tool'*)	**Process** (*'process*)	**Environment** (*'environment'*)
'wing'	**0.7073**	0.2004	0.2594	0.4321	0.1274
'hand'	0.7472 (0.1561)	**0.4466**	0.3578	0.4972 (0.3639)	0.3047
'machine'	0.1912	0.3075	**0.4754**	0.1107	0.1491
'rivet'	0.0000	0.0000	0.0000	**0.3337**	0.0000
'sealing'	0.3655	0.0683	0.2524	**0.3780**	0.0851
'temperature'	0.3510	0.0833	0.1880	**0.6697**	0.3380

Table 1. Lin similarity values for example words, with highest values being highlighted.

5.1 Example

In Figure 4, we illustrate the steps of the proposed method using an example. All the steps were carried out manually (except Part Of Speech tagging), since a computer based implementation is yet to be carried out. The input text is shown, and the location of the assembly issue is indicated using a box. A window of 8 sentences has been selected on either side of the text. For this example we have not considered the document and section titles. A list of nouns is extracted from the text, though it is also possible to extract verbs and other parts of speech. The nouns are compared with the words indicating factors, and classified into one of the five context containers, using similarity measures described above. These labels can then be used as indicators of context while matching context during knowledge application. Note that not all the containers are

uniform in terms of number of words. The classifications are also not perfect, due to reasons such as ambiguity, domain specific terms and noun-phrases.

5.2 Using context

Once these context-containers are available at the stage of application of a piece of knowledge, they can be matched against the situation in which the knowledge is to be applied. Based on the extent to which it matches, a mathematical score can be assigned to indicate the extent of relevance of the piece of knowledge, as shown in Figure 3. In the figure, the matching between the application situation and the context containers would have to be employed in a manner similar to the similarity measures that are used in Table 1.

To illustrate, consider an example based on the knowledge indicated in Figure 4. It is possible that, for an application situation, an assembly operation may involve the process of *gluing*. The ASM for this application situation would contain information about gluing. This would have to be compared with the context containers of the acquired knowledge. Most terms (except for a few for the tool e.g. 'gun' and 'squeeze') are semantically far away from the application situation. However, if the ASM of the application situation had a riveting operation, then the context containers are more likely to be semantically closer, hence more relevant. The similarity between the context containers of acquired knowledge and application situation would not be an exact match, but rather, a number that represents the distance between these two sets of context containers.

6 Future Work

Based on the proposed method discussed in the previous section, the immediate step is to implement it, by capturing the text around the issues, recognizing the entities and events in the surrounding text and other indicators such as the section and document headings. Then, each entity/event can be classified into one of the factors, and attached as a set to each issue in the knowledge base. This is the context of the situation in which the knowledge is acquired. Also, instead of using single words to indicate the assemblability factors, sets of words could be used for each factor (e.g. {'*human*', '*man*', '*ergonomics*', '*person*'} instead of just using '*person*'). Other parts of speech such as verbs, and entities such as noun phrases (e.g. '*riveting gun*') may also be recognized.

Once the context of knowledge is known, it can be utilized during the application of knowledge, using methodologies that have been described in the literature section. Implementations could make use of the context toolkits described, and could also use the ASM to model the application situation. As mentioned in Sec 5.2, the mapping has to be performed by a calculation of similarity between the situation of the acquired knowledge and the application situation.

It was observed in Section 5 that ambiguity of word sense was a factor in comparing similarities. Hence, means for performing word sense disambiguation could be used as a pre-processing step to improve the method. With increase in the size and of number

of texts, use of context vectors, in combination with the current method can also be explored. A further point of exploration is the extent of text surrounding the piece of knowledge to be considered for ascertaining context, potentially leading to application of more large-scale methods such as context vectors and VSMs.

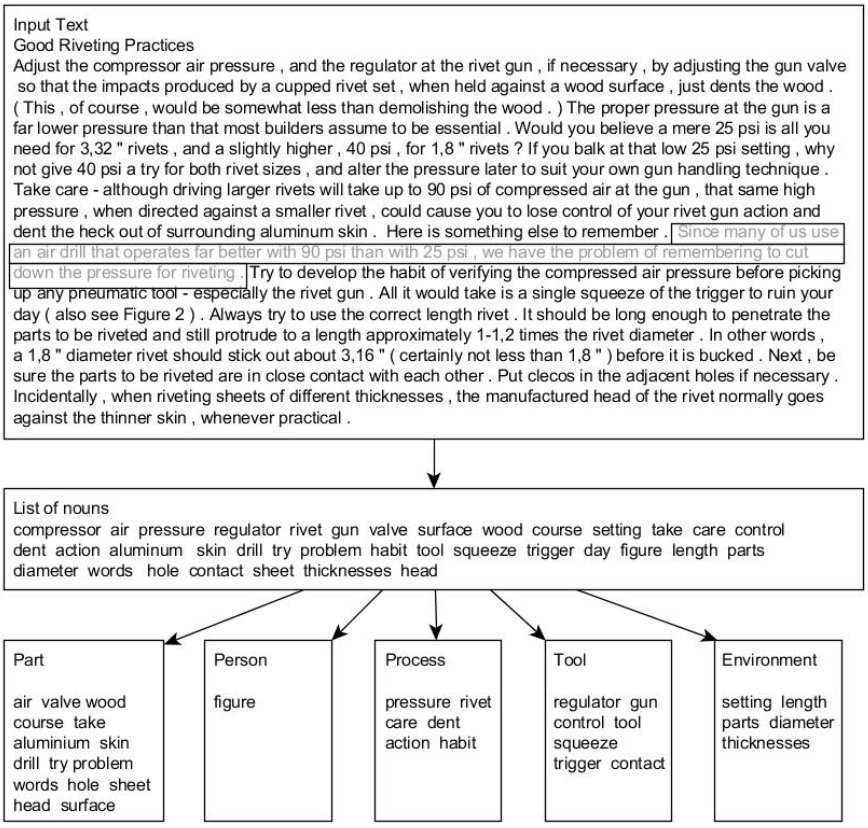

Fig. 4. Example of the working of the proposed method.

7 Conclusions

In this paper, the need for representing context has been explained from the perspective of an ongoing research on acquiring and applying knowledge from documents. The two aspects of assembly knowledge – during acquisition, and during application, have been discussed. Potentially useful indicators, such as an assembly situation model and assemblability factors, have been described for the application and acquisition situations respectively.

The core contribution of this paper is the proposed method for identifying indicators of context during knowledge acquisition. This method uses assemblability factors as

context containers, into which the words surrounding a piece of assembly knowledge can be classified. It is proposed that these words represent the context in which this piece of knowledge was expressed in the document. The next step is to implement this method, and to use the context in the knowledge application stage.

Acknowledgements. The research work described in this paper was carried out with funding from Robert Bosch Centre for Cyber Physical Systems, IISc, under Project RBCO0015.

References

1. Regli, William C., Simon Szykman, and Ram D. Sriram. "The role of knowledge in next-generation product development systems." Journal of computing and information Science in Engineering 1 (2001): 3-11.
2. Ameri, F. and Dutta, D., 2005. Product lifecycle management: closing the knowledge loops. *Computer-Aided Design and Applications*, *2*(5), pp.577-590.
3. Brézillon, Patrick. "Context in human-machine problem solving: A survey." LIP 6, no. 1996 (1996): 029.
4. Henricksen, Karen, Jadwiga Indulska, and Andry Rakotonirainy. "Modeling context information in pervasive computing systems." In International Conference on Pervasive Computing, pp. 167-180. Springer Berlin Heidelberg, 2002.
5. Pomerol, J-C., and Patrick Brézillon. "About some relationships between knowledge and context." In International and Interdisciplinary Conference on Modeling and Using Context, pp. 461-464. Springer Berlin Heidelberg, 2001.
6. Held, Albert, Sven Buchholz, and Alexander Schill. "Modeling of context information for pervasive computing applications." Proceedings of SCI (2002): 167-180.
7. Dey, Anind K., and Gregory D. Abowd. "The context toolkit: Aiding the development of context-aware applications." In Workshop on Software Engineering for wearable and pervasive computing, pp. 431-441. 2000.
8. Bouzeghoub, Amel, Chantal Taconet, Amina Jarraya, Ngoc Kien Do, and Denis Conan. "Complementarity of process-oriented and ontology-based context managers to identify situations." In Digital Information Management (ICDIM), 2010 Fifth International Conference on, pp. 222-229. IEEE, 2010.
9. Patwardhan, Siddharth, and Ted Pedersen. "Using WordNet-based context vectors to estimate the semantic relatedness of concepts." In *Proceedings of the eacl 2006 workshop making sense of sense-bringing computational linguistics and psycholinguistics together*, vol. 1501, pp. 1-8. Trento, 2006.
10. Turney, Peter D., and Patrick Pantel. "From frequency to meaning: Vector space models of semantics." Journal of artificial intelligence research 37 (2010): 141-188.
11. Madhusudanan N., and AmareshChakrabarti. "A model for visualizing mechanical assembly situations." In ICORD 11: Proceedings of the 3rd International Conference on Research into Design Engineering, Bangalore, India, 10.-12.01. 2011. 2011.
12. Santhi, B., Gurumoorthy B and AmareshChakrabarti. "A New approach for Assemblability Assessment Using Time and Postural Analysis—A Case Study." In *ICORD 09: Proceedings of the 2nd International Conference on Research into Design, Bangalore, India 07.-09.01. 2009*. 2009.