A Semantic Similarity Method for Products and Processes

June, 2013

DOCTOR OF ENGINEERING

Suriati Akmal

TOYOHASHI UNIVERSITY OF TECHNOLOGY
Abstract

Due to the complexity of customer requirements coupled with drastic technological changes, development of products and processes is becoming increasingly knowledge intensive. Specifically, retrieving product and process information and making effective use of it requires similarity measures.

Similarity measures are concerned with quantifying of the likeness of the things that are compared. Similarity measures have been practically applied in a wide variety of fields ranging from data mining, case-based reasoning system, image interpretation and pattern recognition. Several researchers have proposed similarity measures that evaluate the likeness between values of numeric properties. However, in many applications some attributes are non-numeric. One solution is to use syntactic similarity measures that calculate the similarity between two words. However, syntactic approaches are limited as they fail to produce good matches when confronted with the meaning associated to the words they compare.

To overcome the above drawbacks semantic similarity measures are been investigated. A semantic similarity measure is a function that quantifies the degree of likeness between two things based on the meaning associated to each thing being compared. This research contributes to the field of semantic similarity measures for products and processes. A novel approach has been proposed in this research, based on Formal Concept Analysis (FCA) and a set of criteria for the characterization of products and processes called Formal Attribute Specification Template (FAST).

This research focuses on countable objects that are represented in terms of their physical aspects and processes in which they are involved. Processes can be intentional or unintentional. In an intentional process, a
particular objective is accomplished. Unintentional processes include natural phenomena and undesired processes such as harmful explosions or fires.

The proposed approach is composed of semantic similarity measures that compare classes in a taxonomy obtained with FCA and a template for the specification of formal attributes (FAST).

The semantic similarity measures of the proposed approach compare classes of products or processes. The comparison is based on the assumption that the more common attributes that are shared by two classes the more similar they are. Therefore, a class is 100% similar to another class if both classes have exactly the same attributes. In particular, the attributes are the formal attributes from the FCA. For this purpose, several similarity equations are investigated in this research by using formal attributes as the sets they compare.

Class taxonomies are defined by means of the subclass relation. A class is a subclass of another class if every member of the subclass is also a member of the super class. Formal Concept Analysis (FCA), which is a method based on applied lattice and order theory, is selected as the taxonomy generator.

FAST helps to describe the formal attributes common to all members of a given class that distinguish them from members of another class. The product formal attributes are expressed in terms of its mereological and topological structure and its involvement with one or more processes. The process formal attributes are expressed in terms of: (1) objects that are always changed by the process (a.k.a inputs); (2) objects that are always produced by the process (a.k.a outputs); (3) participating physical objects (including locations, agents, and performer) other than inputs and outputs; (4) sub-activities that compose the process (a.k.a sub-activities).

The proposed approach was evaluated against edge-counting and information-based similarity measures. In order to quantify the efficacy of each similarity measure, the degree of correlation with human judgment was
used. The results of the evaluation show that the proposed approach performed better than existing similarity measures.

The proposed approach is illustrated with two case studies. The first case study demonstrates the use of FAST for the construction of an ontology for machining processes. The resulting machining processes ontology was evaluated and compared against a third-party ontology. The degree of correlation with Internet-search engine using the value of the Normalized Google distance evaluated the accuracy of each ontology. The results of evaluation show that the ontology obtained with FAST is slightly better than the existing ontology. It was also found that FAST can provide the design rationale of the ontology.

The second case study focused on the application of the proposed semantic similarities for selecting the service strategy for Product-Service systems (PSS) at the early stage of design. It is often the case that the PSS designer is faced with limited amount of knowledge at the early stage of design. One solution is to use the case-based reasoning (CBR) system to facilitate the service strategy selection in which PSS design problems are solved by using or adapting previously obtained design solutions. Existing CBR-systems use numerical similarity measures to search the relevant solution to the problem to be solved. In this case study, a semantic CBR-system was developed by incorporating product-class-comparison based on the proposed semantic similarities. The results of evaluation show that the proposed approach proved useful when some details of information are not available.
Acknowledgement

During my stay in Toyohashi, Japan I was facilitated by many people and I would like to express my gratitude for their support. The first and foremost thanks to God who blessed me with such a great opportunity to learn and granted me success in this long-time endeavor “Alhamdulillah”.

I owe a debt of gratitude to my research advisor, Associate Professor Dr. Rafael Batres, for providing me with a wealth of knowledge, sharing with the times of struggles, and his consistently guiding and inspiring me throughout this endeavor.

I am thankful to Prof. Yoshiaki Shimizu for his valuable input and time as well as accepting and inducting me in Industrial System Engineering Lab. I would like to thank to Prof. Masaki Aono and Associate Prof. Tetsuo Miyake for serving on my doctoral committee. I greatly appreciate for their assistance and input. I also want to thank to all members of Industrial System Engineering Lab for sharing their knowledge, wisdom and inspiration.

I am very obliged to the Ministry of Higher Education of Malaysia and Universiti Teknikal Malaysia Melaka for awarding me with the scholarship to study at Toyohashi University of Technology.

I owe the biggest acknowledgement to my parents, siblings, husband and sons for their understanding, encouragements, patience, love and supports while I am occupied in my research work.
<table>
<thead>
<tr>
<th><strong>Glossary</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case-based reasoning (CBR)</strong></td>
</tr>
<tr>
<td><strong>Edge-based measure</strong></td>
</tr>
<tr>
<td><strong>False positive</strong></td>
</tr>
<tr>
<td><strong>Feature-based measure</strong></td>
</tr>
<tr>
<td><strong>Formal Concept Analysis (FCA)</strong></td>
</tr>
<tr>
<td><strong>Information-based measure</strong></td>
</tr>
<tr>
<td><strong>Mereology</strong></td>
</tr>
<tr>
<td><strong>Mean Absolute Percentage Error (MAPE)</strong></td>
</tr>
<tr>
<td><strong>Ontologies</strong></td>
</tr>
<tr>
<td><strong>OWL</strong></td>
</tr>
<tr>
<td><strong>Process</strong></td>
</tr>
<tr>
<td>Term</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
<tr>
<td>Product</td>
</tr>
<tr>
<td>Product-service system</td>
</tr>
<tr>
<td>Root Mean Squared Error (RMSE)</td>
</tr>
<tr>
<td>Similarity</td>
</tr>
<tr>
<td>Semantic similarity</td>
</tr>
<tr>
<td>Synset</td>
</tr>
<tr>
<td>Topology</td>
</tr>
</tbody>
</table>
Table of Contents

Abstract .................................................................................................................................................. i
Acknowledgement .......................................................................................................................... iv
Glossary ............................................................................................................................................... v
Table of Contents ............................................................................................................................. vii
List of Figures ........................................................................................................................................ x
List of Tables ......................................................................................................................................... xii

INTRODUCTION ........................................................................................................................................ 1
1.1. Similarity measures ....................................................................................................................... 1
1.2. Why are similarity measures necessary for products and processes? ........................................ 6
1.3. Overview of the proposed approach ............................................................................................ 8
1.4. Thesis outline ............................................................................................................................... 10

LITERATURE REVIEW ......................................................................................................................... 12
2.1. Semantic similarity measures .................................................................................................... 12
2.1.1. Edge-based measures ........................................................................................................... 14
2.1.2. Information-based measures .................................................................................................. 16
2.1.3. Feature-based measures ........................................................................................................ 19
2.2. Semantic similarity measures using multiple ontologies ............................................................ 21
2.3. FCA and class hierarchy generation ............................................................................................ 23

THE PROPOSED APPROACH .................................................................................................................. 26
3.1. Introduction .................................................................................................................................. 26
3.2. Product representation ............................................................................................................... 32
3.3. Process representation ............................................................................................................... 34
3.4. Formal Attribute Specification Template (FAST) ........................................................................ 35
3.5. Procedure for taxonomy construction with FCA ........................................................................ 38
3.6. The semantic similarity measures of the proposed approach ..................................................... 40

EVALUATION OF THE PROPOSED APPROACH .................................................................. 48
4.1. A taxonomy for home electric appliances ...................................................................................... 48
4.1.1. Taxonomy construction ................................................................. 49
4.2. Evaluation of the proposed approach ............................................... 53
4.2.1. Similarity calculation ................................................................. 55
4.2.2. Experiment results ................................................................. 55
4.2.3. Analysis of the results ............................................................. 56
4.3. Conclusions .................................................................................. 59

APPLICATION OF THE PROPOSED APPROACH TO PROCESS ONTOLOGY DEVELOPMENT ................................................. 62
5.1. An ontology for machining processes ............................................. 62
5.1.1. Ontology construction ............................................................. 64
5.1.2. Ontology evaluation and remediation ....................................... 75
5.2. Web-based Evaluation of Semantic Similarity Measures ............... 79
5.3. Conclusions .................................................................................. 87

APPLICATION OF THE PROPOSED APPROACH FOR PRODUCT-SERVICE SYSTEMS DEVELOPMENT ................................................. 88
6.1. Introduction ................................................................................... 88
6.2. Selection of services in product service-systems ............................ 90
6.3. Methodology for product service-systems based on semantic similarity 92
6.4. Ontology construction ................................................................. 93
6.5. Case study .................................................................................. 96
6.6. Conclusions .................................................................................. 111

CONCLUSIONS AND FUTURE WORK ......................................................... 113
7.1. Major contributions ................................................................. 113
7.2. Future work ............................................................................... 115

References .......................................................................................... 117

Appendix A: Formal Concept Analysis ............................................... 124
Appendix B: Draw a concept lattice by hand ....................................... 128
Appendix C: Questionnaire ............................................................... 131
Appendix D: Existing methodologies for constructing ontologies .......... 134
Appendix E: Ontologies applications in products and processes .......... 136
Appendix F: RMSE, MAPE, R for machining process ontologies .......... 138
Appendix G: Regression analysis ......................................................... 141
Appendix H: Successful PSS cases and case description .................. 144
Appendix I: Precision-recall graphs .................................................... 147
List of Figures

Fig. 1-1 The function of a product are the desired behavior of a product...........2
Fig. 3-1 Activity representation in IDEF0..............................................................28
Fig. 3-2 Activity in Upper ontology based on ISO 15926................................. 29
Fig. 3-3 Composition of device and its relation to processes.....................32
Fig. 3-4 Composition of device presentation....................................................33
Fig. 3-5 Flow diagram for the formal attribute selection of a given class of product........................................................................................................36
Fig. 3-6 Flow diagram for the attribute identification of a given candidate class........................................................................................................37
Fig. 3-7 A sample for similarity calculation.......................................................44
Fig. 4-1 Context table for an ontology of home electric appliances ............51
Fig. 4-2 Concept lattice obtained with the context table of Fig. 4-1..............52
Fig. 4-3 The level of inconsistency for each questionnaire..........................54
Fig. 4-4 Terms pair integrity................................................................................54
Fig. 4-5 Results of the cluster analysis...............................................................59
Fig. 5-1 Lattice obtained with the preliminary context table ......................70
Fig. 5-2 (a) boring enlarges a hole; (b) reaming produces a slightly enlarged a hole that has a more accurate diameter; (c) counterboring enlarges a part of the hole so that the bottom part of the enlarged portion of the hole is flat and square...........................................................................................................71
Fig. 5-3 The revised concept lattice.................................................................73
Fig. 5-4 Class hierarchy of machining processes............................................74
Fig. 5-5 Regression analysis for Wu Palmer similarity......................................83
Fig. 5-6 The repositioning of class tapping in machining ontology............86
Fig. 6-1 Methodology for PSS ........................................................................94
Fig. 6-2 The concept lattice for PSS..............................................................95
Fig. 6-3 Class hierarchy for product ontology ...............................................96
Fig. 6-4 Screen dump of the user interface of the semantic CBR system....105
Fig. 6-5 Precision (upper part) – recall (lower part) graphs for problem No. 37 ................................................................. 109
Fig. 6-6 Precision (upper part) – recall (lower part) graphs for problem No. 43 ................................................................. 110
List of Tables

Table 3-1  Comparison of process representations.................................................. 31
Table 3-2 Association among sets of attributes of classes being compared... 46
Table 4-1 Comparison between similarity measures.................................................. 58
Table 5-1  List of potential classes and formal attributes for machining processes................................................................................................................................. 67
Table 5-2 Preliminary context table........................................................................... 68
Table 5-3 Modified context table.............................................................................. 72
Table 5-4 Evaluation of $C_1$=counterboring using the class hierarchy of the machining ontology .......................................................................................................................... 77
Table 5-5 Performance of each class in the machining ontology ........................ 78
Table 5-6 Average, minimum, and maximum values of $RMSE$, $MAPE$ and $R_..$ 78
Table 5-7 The results for experiments of material removal process ................. 81
Table 5-8 The studentized residual for each similarity measure in comparison against the Google Scholar ............................................................................................................................ 84
Table 5-9 The correlation coefficient, sum of squared errors and standard deviation for machining ontology .......................................................................................................................... 85
Table 6-1 The results for experiments of product-service system.................... 100
Table 6-2 Comparison of identical cases ............................................................. 106
Table 6-3 Comparison of average precision and recall..................................... 111
INTRODUCTION

1.1. Similarity measures

Generally, “similarity” is a term to enclose whether two things, or two situations are similar or dissimilar. According to [1], similarity plays an important role in studies of theories of cognition and how people make comparisons. According to [2], “similarity is a core element for learning, knowledge and thought, for only our sense of similarity allows us to order things into kinds so that these can function as stimulus meanings reasonable expectation depends on the similarity of circumstances and on our tendency to expect that similar cause will have similar effects”.

According to Holt [3], similarity is important for humans to understand the existence of objects, structure and actions together with their connections in reality. The degree to which we determine if two things are similar is both intuitive and based on our knowledge. For example, when an individual plans to use a toaster on the dining table as shown in Fig. 1-1, he or she will imagine the
result of using the toaster, which is related to the function performed by the product. The memory, which has some prior knowledge, organizes the information and somehow translates it into associations such as bread and toaster, toasted bread and toaster. Based on memory of the past, a toaster is always used to toast the bread. When comparing a toaster and let’s say a pizza oven, we are inclined to look at common aspects such as the use of heat to produce warm and somehow crispy bread.

Fig. 1-1 The function of a product are the desired behavior of a product

In addition, if only a few objects are given, it is easy for a human to identify how close two objects are by finding their common aspects. However, it becomes more complex for a large numbers of objects.

Therefore many practical applications require computational similarity measures. As a matter of fact, the computational approaches for measuring similarity that emphasize imitate the
human ability of assessing similarity between two things date back to [4].

The past decade has seen the development of computational similarity measure that are based on geometric models that assume objects are represented by points in some coordinate space. The similarity of these approaches is calculated by the metric distance between respective points. However, one of major problems with this approach is the inappropriateness to represent the dimensional representation for qualitative properties of thing being compared [5].

In recent years, there has been an increasing interest in using feature comparison to quantify the degree of likeness of the things that are compared. Tversky and Gati [5] identified similarity as a function that quantifies the degree to which two sets of features match each other. They proposed a similarity that considers both common and distinct features which are known as the contrast model. Their contrast model explained that the similarity should not be viewed as a symmetric relation such as a is similar to b than b is similar to a. For example, people say “the son resembles the father” rather than “the father resembles the son”; “the portrait resembles the person” and not “the person resembles the portrait”. Russel and Norvig [6] defined similarity as an evaluation of the common intrinsic features shared by two things. The intrinsic features are the important features that belong to a thing. If the thing is described without this feature, the meaning of the thing is incomplete.

Similarity measures play an important role in information retrieval process, information extraction, information integration and other applications involving comparison two things. In an information retrieval system, determining the optimal match between a queries and stored information is the fundamental
operation that highly depends on similarity measures. In such systems, the retrieved information is sorted in order of their decreasing similarity. High-ranked information is likely to have similar properties to the query.

Also, similarity measures can be used for problem solving. For example, the case-based reasoning systems use reasoning that draw conclusion by similarity. It imitates human reasoning for solving a problem by making use of the previous experiences.

Similarity measures in pattern recognition are used for classifying sets of objects into classes. Similar objects are grouped within the same cluster and dissimilar objects in different cluster.

In numerous multimedia processing systems and applications, assessment of image similarity is important for image copy detection, retrieval and recognition problem. Similarity measures are used to interpret the characteristics of an image that compared against its variations versions such as contrast/brightness variation.

Although numerous concept of similarity measures have been applied in many scientific fields and presented in many forms and interpretations, they all have in common of comparing two objects, two situations, for various reasons including knowledge, biases and goals [7].

Most similarity measures evaluate differences between values of numeric attributes such as in the numerical difference between two given diameter values. However, many applications require non-numeric similarities as well. For example, case-based reasoning systems for the conceptual design of products and processes must be developed to work with a limited knowledge about the products and processes.
Nearly all of non-numeric similarity measures are based on syntactic grounds. For example, the Levenshtein distance \cite{16}, \cite{17} can be used to calculate the similarity between two words, in terms of the minimum number of operations that are needed to transform one of the words into the other. However, from the point of view of the meaning of the words that are compared, existing syntactic similarity-measures often result in incorrect matches.

Semantic similarity measures can be used in order to overcome the limitations of syntactic approaches. A semantic similarity is a function that assigns a numeric value to the similarity between two classes of objects based on the meaning associated to each of the objects \cite{18}. For a review of semantic similarity metrics, the reader is referred to the paper of Cross and Hu \cite{19}.

Recently, the use of ontologies for evaluating similarity has been reported in the literature \cite{20}, \cite{21}. Ontologies are formal models that use mathematical logic to disambiguate and define classes of things \cite{22}. Specifically, ontologies describe a shared and common understanding of a domain in terms of classes, possible relations between things, and axioms that constrain the meaning of classes and relations \cite{23}. A class represents a set of things that share the same attributes. A relation is used to represent a relationship among two or more things. Examples of relations are less than, connected to, and part of. Class taxonomies are defined by means of the subclass relation. A class is a subclass of another class if every member of the subclass is also a member of the super class. Axioms are typically represented as logic constructions that formally define a given class or relation.

Most semantic similarities are defined in terms of the number of edges between the classes that they compare. The research to date
has tended to focus semantic similarities that are defined in terms of features but uses synsets for the comparison between words rather than classes. Most of the existing similarities measures use a large database such as WordNet for general purpose and Mesh for medical purpose for evaluating the word comparison.

In this thesis, a comprehensive approach towards the similarity measures for products and processes information that can deal with non-attribute information is developed.

1.2. Why are similarity measures necessary for products and processes?

A product is defined as something that is the result of a process. On the other hand, typical chemical engineering textbooks define a process as “an operation or a series of operations” that “cause a physical or chemical change in a substance or mixture of substances” [8]. Textbooks also explain that processes commonly have several steps, each of which represents a specific physical or chemical change. Such definitions assume that during the realization of a process, a particular objective is accomplished. In other words, according to these definitions, a process has a design intention.

However, unintentional phenomena are also of concern to engineers. For example, explosions (such as those that result in property damage) may happen as a result of an abnormal situation rather than a well-designed series of steps. Despite differences related to whether an objective is involved or not, both intentional and unintentional processes share the ability to transform material or energy through one or more changes. This research addresses both kinds of processes.
Due to the complexity of customer requirements coupled with drastic technological changes, development of products and processes is becoming increasingly knowledge intensive. It brings about change in the way industries organize products and processes. The market demands industries to effectively manage the know-how about products and processes as a means to differentiate the business competitions. Information about products and processes has to be considered as a rather special resource [9]: it does not get lost when it is used, and the costs for generating and procuring information are high compared to the costs for its storage and dissemination.

Product development which is a multi-disciplinary in nature requires a variety of product life-cycle knowledge [10]. Specifically, design teams face a considerable challenge in making effective use of increasing amounts of information that is stored in several information systems. Also, it is often the case that product designers can reuse past designs rather than designing from scratch [11]. Thus it would be very important to have the ability to retrieve product data.

As mentioned above, information retrieval consists of translating and matching a query against a set of information objects. The information retrieval system responds to the query using a given algorithm and a similarity measure. Particularly, information retrieval plays an important role in areas such as product family design [12], product embodiment, and detailed design [13]. Shah et al. [14] present a combination framework that consists of software engineering, data engineering and knowledge engineering and design theory.
In order to support product and process information retrieval and reuse, some authors suggest the use of case-based reasoning (CBR) in which design problems are solved by using or adapting previous design solutions [13], [15].

A CBR system is composed of domain knowledge, a case base and a search mechanism based on a similarity measure. Domain knowledge refers to knowledge about the features of the different objects or entities that a case is about. A case base contains a set of cases, each of which describes a problem and a solution to the problem. The problem is typically defined in terms of specific features of objects. Finally, a similarity measure quantifies the differences that exist between objects [7]. CBR uses similarity measures to identify cases which are more relevant to the problem to be solved.

1.3. Overview of the proposed approach

The objective of this thesis is to develop a more effective semantic similarity method for products and processes. The proposed approach is composed of semantic similarity measures that compare classes in a taxonomy obtained with Formal Concept Analysis (FCA) and a template for the specification of formal attributes.

The proposed approach is based on two main pillars. One is a semantic similarity measure based on Formal Concept Analysis (FCA). The semantic similarity measure of the proposed approach compare classes of products and processes. The semantic similarity measure is emphasized on the common formal attributes that are obtained from FCA. It is a method based on applied lattice and order
theory, is selected as the taxonomy generator. The underlying principle in this research is that if a class represents a set of things that share the same attributes (such as a class in a taxonomy), we can state that a class is equivalent to another class if both classes have exactly the same attributes. This implies that the more common attributes that are shared by two classes the more similar they are. For this purpose, several similarity equations are investigated in this research by using formal attributes as the sets they compare. It became clear that the sets of features could be replaced with sets of formal attributes from the FCA.

The second pillar is a new way to specify the formal attributes required by FCA. This method is referred to as Formal Attribute Specification Template (FAST). FAST identifies the product formal attributes by considering its mereological and topological structure and its involvement with one or more processes. FAST also identifies the formal attributes of processes.

The proposed semantic similarity method consists of two steps: taxonomy generation and similarity calculation.

FAST is used in the taxonomy generation for formal attribute identification which is later used in FCA to generate a lattice. The resulting lattice and formal attribute information obtained with FCA are later used to create a class hierarchy.

In the second step, similarity between two classes of this taxonomy is calculated using a semantic similarity measure, in which the taxonomy structure and formal attribute information are used as input. For this purpose, the edge-counting and information-based similarity measures were used to evaluate and compare against the proposed approach. In order to quantify the efficacy of each similarity measure, the degree of correlation with human
judgment and NGD similarity were used. The results of the evaluation show that the proposed approach performed better than existing similarity measures.

1.4. Thesis outline

The remainder of this thesis consists of six chapters followed by bibliography. Topics discussed in every chapter are as follows:
Chapter 2 contains a comprehensive description on concept of semantic similarity and presents an overview of common semantic similarity measures.

Chapter 3 describes the contribution of this research. This chapter introduces the semantic similarity equation and the Formal Attribute Specification Template (FAST) of the proposed approach.

In chapter 4, the proposed approach is evaluated and compared against the existing similarity measures. The correlation of each similarity score is compared against the human similarity ratings.

Chapter 5 describes the application of the proposed approach for constructing machining process ontology. The resulted machining ontology was evaluated and compared against a third-party ontology. The degree of correlation with Internet-search engine using the value of the Normalized Google distance evaluated the accuracy of each ontology.

Chapter 6 demonstrates a real-world application in product-service system. In this research, the existing CBR systems that use numerical