WMG Innovative Solutions

WMG Service Systems Research Group
Working Paper Series

Using Ontology Engineering for Understanding Needs and Allocating Resources in Web-Based Industrial Virtual Collaboration Systems

---

Xiao Ma
Jay Bal
Ahmad Issa

THE UNIVERSITY OF WARWICK

ISSN: 2049-4297
Issue Number: 01/14

About WMG Service Systems Group

The Service Systems research group at WMG works in collaboration with large organisations such as GlaxoSmithKline, Rolls-Royce, BAE Systems, IBM, Ministry of Defence as well as with SMEs researching into value constellations, new business models and value-creating service systems of people, product, service and technology.

The group conducts research that is capable of solving real problems in practice (ie. how and what do do), while also understanding theoretical abstractions from research (ie. why) so that the knowledge results in high-level publications necessary for its transfer across sector and industry. This approach ensures that the knowledge we create is relevant, impactful and grounded in research.

In particular, we pursue the knowledge of service systems for value co-creation that is replicable, scalable and transferable so that we can address some of the most difficult challenges faced by businesses, markets and society.

Research Streams

The WMG Service Systems research group conducts research that is capable of solving real problems in practice, and also to create theoretical abstractions from or research that is relevant and applicable across sector and industry, so that the impact of our research is substantial.

The group currently conducts research under six broad themes:

- Contextualisation
- Dematerialisation
- Service Design
- Value and Business Models
- Visualisation
- Viable Service Systems and Transformation

WMG Service Systems Research Group Working Paper Series
Issue number: 01/14
ISSN: 2049-4297
January 2014

Using Ontology Engineering for Understanding Needs and Allocating Resources in Web-Based Industrial Virtual Collaboration Systems

Xiao Ma
Senior Research Fellow
Service Systems Group, Warwick Manufacturing Group,
University of Warwick, Coventry CV4 7AL, UK.
Tel: +44 (0) 2476524718. E-mail: X.Ma@warwick.ac.uk

Jay Bal
Associate Professor
Warwick Manufacturing Group
University of Warwick, Coventry CV4 7AL, UK.
Tel: +44 (0) 24765 24290. E-mail: Jay.Bal@warwick.ac.uk

Ahmad Issa
Doctoral Candidate
Warwick Manufacturing Group
University of Warwick, Coventry CV4 7AL, UK.
Tel: +44 (0) 2476524718. E-mail: a.issa@warwick.ac.uk

Acknowledgement: This work was funded by the Engineering and Physical Sciences Research Council (EPSRC) via the Warwick Innovative Manufacturing Research Centre (WIMRC) to investigate ontology driven industry collaborative platforms for improving processes.

If you wish to cite this paper, please use the following reference:

Ma X, Bal, J, & Issa A (2014) Using Ontology Engineering for Understanding Needs and Allocating Resources in Web-Based Industrial Virtual Collaboration Systems, *Computers in Industry, forthcoming.* Interim location: *WMG Service Systems Research Group Working Paper Series*, paper number 01/14, ISSN 2049-4297.

Using Ontology Engineering for Understanding Needs and Allocating Resources in Web-Based Industrial Virtual Collaboration Systems

Abstract

In many interactions in cross-industrial and inter-industrial collaboration, analysis and understanding of relative specialist and non-specialist language is one of the most pressing challenges when trying to build multi-party, multi-disciplinary collaboration system. Hence, identifying the scope of the language used and then understanding the relationships between the language entities are key problems. In computer science, ontologies are used to provide a common vocabulary for a domain of interest together with descriptions of the meaning of terms and relationships between them, like in an encyclopedia. These, however, often lack the fuzziness required for human orientated systems. This paper uses an engineering sector business collaboration system (**www.wmccm.co.uk**) as a case study to illustrate the issues.

The purpose of this paper is to introduce a novel ontology engineering methodology, which generates structurally enriched cross domain ontologies economically, quickly and reliably. A semantic relationship analysis of the Google Search Engine Index was devised and evaluated. Using Semantic analysis seems to generate a viable list of subject terms. A social network analysis of the semantically derived terms was conducted to generate a decision support network with rich relationships between terms. The derived ontology was quicker to generate, provided richer internal relationships and relied far less on expert contribution. More importantly, it improved the collaboration matching capability of WMCCM.

Keywords: Ontology Engineering, Self-help Systems, Semantic Web, Semantic Relationship, Social Network Analysis

1. Introduction

The increasing need for information exchange has driven the interest in ontology generation [1, 2], and engineering was among the earliest sectors to benefit. Ontologies in this sector are considered to be more mature than in other such sectors. Ontologies in this sense are increasingly used in knowledge management systems, medical and bio-informatics and are set to play a key role in the semantic web and grid computing.

In this research the requirement for an effective ontology system came from the West Midlands Collaborative Commerce Market Place (WMCCM). This is a web portal matching "need" with "competence" and enabling collaborations among SMEs to address overall tender needs through a combination of competencies [3]. In order to automate the matching process between companies and tenders, WMCCM classifies company competencies against a three level ontology. It also semantically analyses every incoming tender to identify what competencies are required and maps these onto the same ontology. This allows WMCCM system to forward tenders to companies that have the right capability, or to form partnerships.

A key factor affecting the effectiveness of the matching functions is the quality of the ontology that links tenders with company capability. The WMCCM engineering ontology was built in a fairly orthodox way, the re-use of previously published ontology and adaptation/modification by experts. Thus it followed a mixed approach: lower levels were derived from actual company interview information; upper levels from standard classifications such as the United Nations Standard Products and Services Code (UNSPSC) and Standard Industry Classification (SIC).

UNSPSC was designed as an upper level ontology to facilitate e-Business for quicker and more accurate procurement, marketing and sales. It was designed for high level guidance, and it does not appear to be practical at the regional and country level [4].

For example, The United Kingdom Standard Industrial Classification of Economic Activities (UK SIC) is the standard industrial classification widely accepted in the UK. It is used to categorise businesses in accordance with the scope of their economic activity[5].

Although fundamentally UNSPSC and SIC were supposed to represent the same knowledge and its structure, UNSPSC lacks domain coverage, especially with regard to actual products and services, and there are insufficient relationships to provide inheritance and commonality among classes[6]. This illustrates that while many ontology have reused such sources, they still require considerable consultancy from domain experts to clarify the relationships between such sources. [7].

These issues suggest that directly summarising ontology from existing sources will not satisfy WMCCM's requirement for broad coverage and rich internal relationship. Therefore, WMCCM followed a mixture of top down derivation and bottom up synthesis collecting terms and relationships from actual business users.

However, this customisation did not fully satisfy WMCCM's tender matching process. The source ontology (UNSPSC and SIC) lacked the necessary level of fuzziness/redundancy to be able to be applied to human oriented systems. Consequently, the reuse of such ontology only provides the necessary structure and description of domain knowledge, but lack relationships to terms that are not strictly bounded by the core domain terms. The required fuzziness may be gained by increasing the number of semantic relationships with terminology which is not exclusive to engineering domain. The "relationship sea" with rich internal relationships among concepts needs to be expanded in order to contain a network of mutually inclusive terms for multi-disciplinary usage.

Therefore, this work set out to address these issues and describes a novel methodology which generates ontology for a specific domain economically, quickly and reliably, and builds a rich relationship sea of semantically related domain terms.

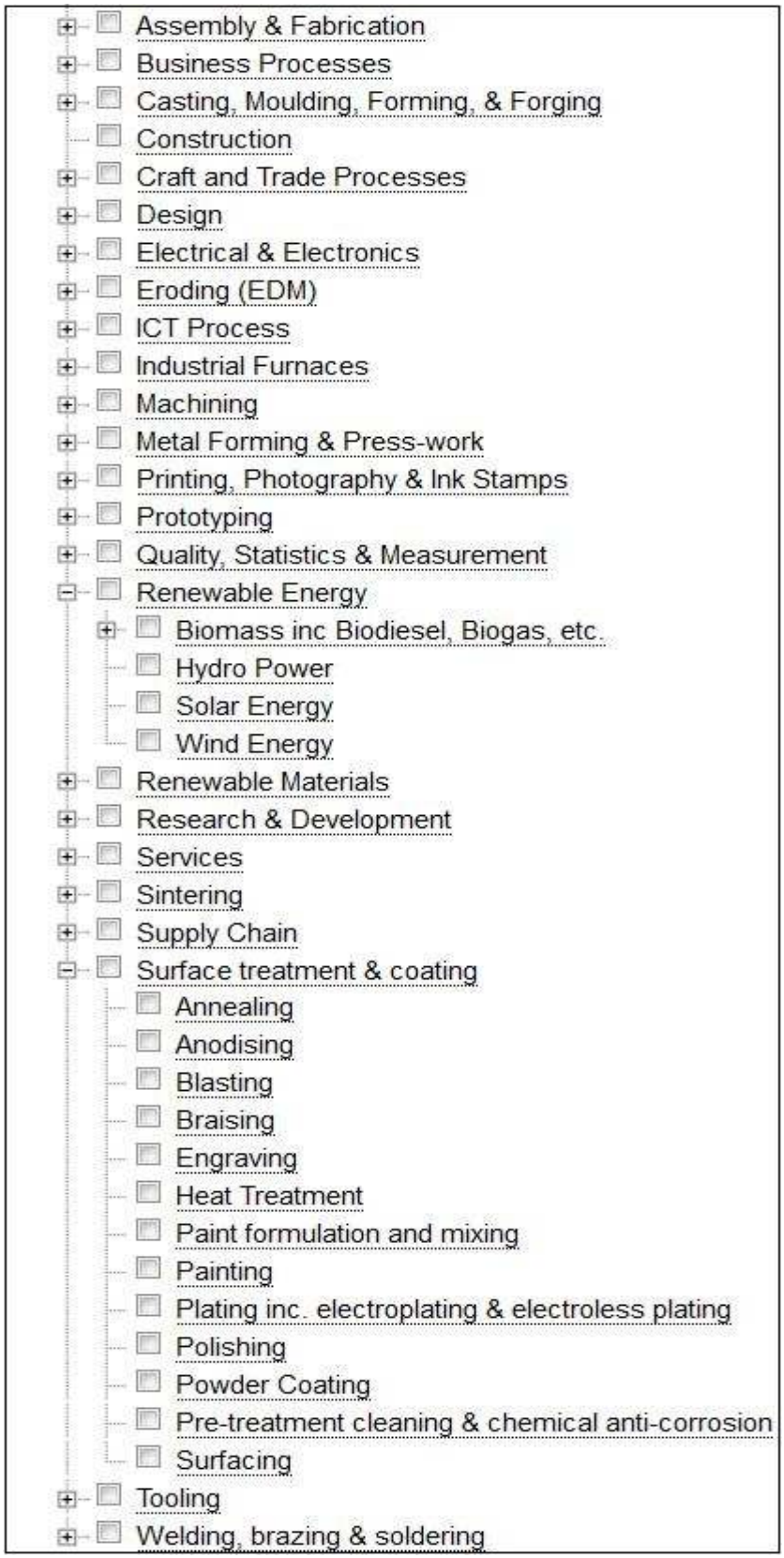

Figure 1: Current WMCCM Ontology[1]

[1] It is a three level tree structure, where only the "Renewable Energy" and "Surface Treatment & Coating" sections are expanded in this figure.

2. Keyword Grouping

The first step of building the ontology is to identify the terms within the target domain. The terms are group(s) of concepts representing similar domain concepts to the seeding words. Techniques which provide grouped domain concepts are:

- Categorisation: “a method provides groups of entities whose members are, in some way, similar to each other”. It concentrates on “concept formation and coverage” and allows overlapping [7]
- Classification (including taxonomy): “a method involves the orderly and systematic assignment of each entity to one class within a system”. It highlights “only one class and no overlapping”[7], and emphasizes “delimiting and distinguish”[8].

Categorisation better meets the research purpose, as allowing overlap can create keywords groups to maximize coverage over target subject areas. Focusing on overlapping coverage allows fuzzy concepts which link the terminology in the concept to other concepts in the domain or to other domains and also importantly to the non-specialist language in a domain.

Within categorisation techniques, a method called “Word Clustering” directly utilises “co-appearing in content” forming the semantic relationship between terms. Two different types of word similarity have been used in word clustering:

- Semantic similarity: two words that are paradigmatically similar (thesaurus), and substitutable in a particular context. For example, “I ate sausages for breakfast”, the word sausages can be substituted by “bacon” with little change to the meaning and structure of the sentence, and therefore these two words can be identified as being semantically similar;
- Semantic relatedness: two words that often occur simultaneously in a text. For instance, fire and burn are semantically related, since they often appear together within the same context[9].

This research focuses on semantic relatedness rather than semantic similarity. This is because keywords representing the same concept are more likely to co-occur in sentences, but are not necessarily substitutable with each other.

3. Research Methodology

The used methodology for building the ontology is based on the principle that the ontology building should be initialised by linking specified keywords to the target source. SENSUS (Swartout et al., 1997) constructs ontology for a domain from the foundation of a large knowledge base, or ideally, a previous large ontology. However, it does not engage in a traditional reusing or re-engineering process. It identifies key domain specific terms, a.k.a. seeding words, and then links them to the large ontology. Afterwards, the terms irrelevant to the new ontology can be pruned from the large source ontology. The following processes should be undertaken in the SENSUS approach (Figure 2):

Figure 2: Approach to developing a SENSUS ontology

This approach contains unique characterises that provide advantages over the other approaches:

- It is an obvious improvement that SENSUS no longer requires constant input from domain experts: it only needs the initial seeding terms and their relationships to the knowledge base.
- SENSUS combines corpus construction and ontological analysis in one process instead of keep them separate [12 - Cyc][13 - Methontology]. Therefore, SENSUS ensures the terms collected were semantically connected to the seeding terms.
- The development of different ontology shared the same knowledge bases and their internal links. One of the main advantages of SENSUS was that the massive coverage of the SENSUS ontology becomes a "hinge" that marries the terminology and the organization of other ontology developed that are based on it (Swartout et al., 1997).
- Extracting related terms from the same sources by different seeding words is similar to perceiving the same knowledge from different perspectives. This in theory could result in fuzziness around any given concept. Thus the SENSUS ontology construction method may be capable of building cross-domain ontology.

The SENSUS methodology seems superior to the others in the ways discussed. However, it is difficult to reuse SENSUS directly, as there is insufficient detail on the techniques suggested. In addition, SENSUS did not propose any post-development stage, a development life cycle or project management mechanism. Therefore, this research used the SENSUS approach as a foundation approach and developed techniques to formulate a new methodology that met the needs for faster, more economical, reliable, multi-domain ontology construction.

3.1. Data Source Selection

Word clustering was chosen as the method to generate keywords to describe structure around a given concept (will be called 'keywords set' hereafter). There are basically two main data sources (corpus) that could be used to generate these keywords:

1. Directly collected expert and user data: first hand data;
2. Directly reused or extracted data from existing data sources which contains words with either their semantically similar or semantically related relationships. There are five types of such sources:
   a. Thesaurus or Dictionaries;

b. WordNet;
c. Industry/Government Codes;
d. Search Engine Index.

The research requirement for less reliance on domain experts, broader coverage of concepts and richer internal relationships means that the use of first hand data is not suitable since it requires significant input from domain experts. In addition, the use of semantic relatedness means that thesaurus/dictionaries and WordNet are not suitable source knowledge bases. Thus a general search engine index, which crawls all types of web pages on the Internet, may better suit the need of this research for a broad coverage, latest developments and rich relationships.

There are many popular search engines available across the Internet, such as Google, Yahoo, and Bing. Among these, Google has been widely regarded as the leader with the largest indexed content and popularity [10, 11]. Uniquely, Google provides a method – Google Sets [12] – to generate "on-topic" terms based on given examples. This method seems to provide an opportunity to generate domain related terms with wider but not chaotic relationships.

3.2. Seeding Words Configuration

The Google Sets tool could link the seeding words to the Google index via semantic relationships, since it is a word clustering tool which extracts semantically associated words from the Google index. Google Sets (Figure 3) has several parameters that can be altered through the Google Sets settings, and the effects of varying these on the semantically related words generated were not yet clear. This required a study of the Google Sets parameters so that they could be configured to provide the best results.

Figure 3: The Google Sets platform

Early experiments to test the quantity and quality of predictions showed that paired keywords generated much better results than any other option. Paired seeding words had the advantage of producing a more focused domain terms, and it seems that paired seeding words particularly benefitted the domain description density for both less focused domains and more naturally focused domains. Therefore paired seeding words were utilised for generating the engineering ontology.

However, a further issue was the need to avoid seeding words that had high potential for misleading the search direction. Therefore, further experiments were conducted to identify the minimum number of required seeding word pairs required to provide reasonable fault tolerance. The results showed that two pairs of keywords appear to be the minimum required. However, two pairs of seeding words may produce predictions around two subject areas. In an extreme case (Figure 3), if a pair did not produce any target domain prediction at all, the experiment may end up with two separate distributions of terms, with no overlap. In such a case, the resulting corpus of terms may not target any particular domain, and further expert guidance may be required. Using three pairs, the system will better tolerate poor seeding word choices, and ensure the output is more reliable.

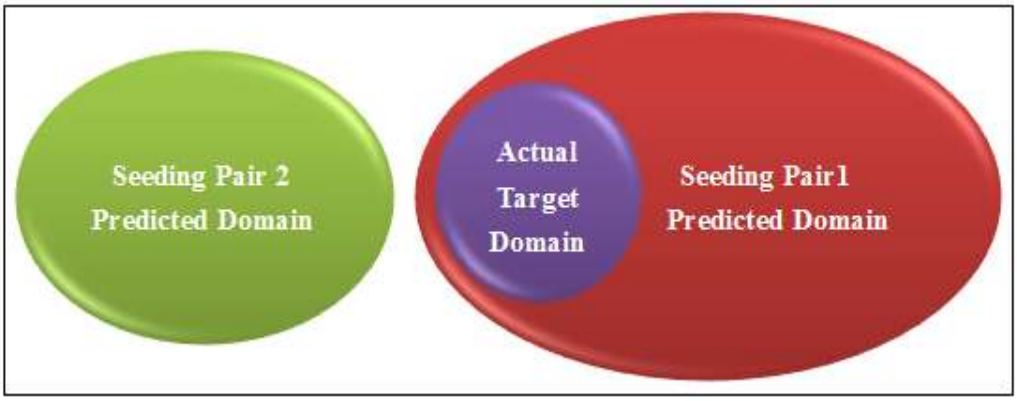

Figure 4: Complete Prediction Separation of Two Pairs of Seeding Words

3.3. Seeding Words Selection

Seeding words for this research were produced from both ontology builders and domain experts. It was expected ontology builders could contribute from application specification of terms, and the domain experts may strengthen the terms' domain representativeness in general. A Delphi approach to collect seeding words for a subject area from domain experts was adopted[13]. This method collects the opinions of different individuals, in order to increase the opportunity of picking objective seeding words and minimize subjective bias from direct study of the application environment.

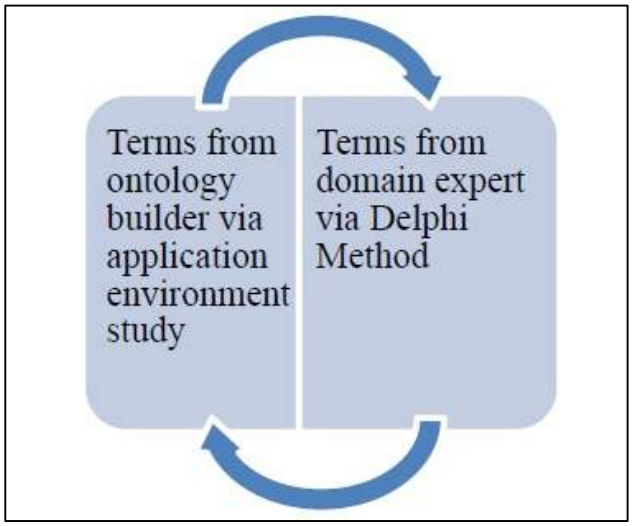

Figure 5: Illustration of seeding words selection

3.4. Corpus Construction

Google Sets was used to generate semantically related terms from the initial seeding words. However, the resulting terms were too few to represent any practical domain or to yield any statistically relevant results. To generate more keywords, the resulting terms were re-input as seeding words again to obtain yet more predicted terms. After this second round of seeding there was better coverage of the domain, but still insufficient concepts and relationships to yield any statistical reliability. Therefore the terms

generated from the second round were used as seeding words to derive third level predictions.

This method is known as "Snowball Sampling" and is common in social studies and statistics, especially within social network analysis[14]. This approach generates a large collection of related entities to construct complex social networks[15]. There are associated social network analysis techniques to uncover more facts about such a network.

In the applied methodology, k1&k2, k3&k4, k5&k6 are defined as three pairs of keywords selected for a chosen domain/application *M* (where *M* is the concept/definition of the domain(s)). These keywords are usually supplied by domain experts, or maybe taken from an existing ontology.

Function $f_{GS}(x,y)$ is the process to capture Google Sets predicts by using given paired seeding keywords x and y. Set $S_{(x,y)}$ represents the collection of the predicted keywords, from $k_1^{x,y}$ to $k_n^{x,y}$ which were generated by function $f_{GS}(x,y)$.

$$S_{(k1,k2)} = f_{GS}(k1,k2) = \left\{ k_1^{1,2}, k_2^{1,2}, \ldots, k_{(n_{1,2}-1)}^{1,2}, k_{n_{1,2}}^{1,2} \right\}$$

Then, in order to generate more optimised outputs, the second round collects the predictions from the first round and pairs them up with the original seeding words as new seeding pairs, and then obtains the new extended predictions from Google Sets. Extended collection for k_1&k_2:

$$S_{\left(k_1,k_1^{1,2}\right)} = f_{GS}\left(k_1,k_1^{1,2}\right) = \left\{ k_1^{1,1,1,2}, k_2^{1,1,1,2}, \ldots, k_{(n_{1,1,1,2}-1)}^{1,1,1,2}, k_{n_{1,1,1,2}}^{1,1,1,2} \right\}$$

$$\vdots$$

$$n_{1,2}$$

$$\vdots$$

$$S_{\left(k_1,k_{n_{1,2}}^{1,2}\right)} = f_{GS}\left(k_1,k_{n_{1,2}}^{1,2}\right) = \left\{ k_1^{1,(n_{1,2}),1,2}, k_2^{1,(n_{1,2}),1,2}, \ldots, k_{(n_{1,(n_{1,2}),1,2}-1)}^{1,(n_{1,2}),1,2}, k_{n_{1,(n_{1,2}),1,2}}^{1,(n_{1,2}),1,2} \right\}$$

$$S_{\left(k_2,k_1^{1,2}\right)} = f_{GS}\left(k_2,k_1^{1,2}\right) = \left\{ k_1^{2,1,1,2}, k_2^{2,1,1,2}, \ldots, k_{(n_{2,1,1,2}-1)}^{2,1,1,2}, k_{n_{2,1,1,2}}^{2,1,1,2} \right\}$$

$$\vdots$$

$$n_{1,2}$$

$$\vdots$$

$$S_{\left(k_2,k_{n_{1,2}}^{1,2}\right)} = f_{GS}\left(k_2,k_{n_{1,2}}^{1,2}\right) = \left\{ k_1^{2,(n_{1,2}),1,2}, k_2^{2,(n_{1,2}),1,2}, \ldots, k_{(n_{2,(n_{1,2}),1,2}-1)}^{2,(n_{1,2}),1,2}, k_{n_{2,(n_{1,2}),1,2}}^{2,(n_{1,2}),1,2} \right\}$$

The same formula is applied to the rest of the first round predictions. Then "snowballing" to get wider domain(s) coverage, all the unique predictions from the second round (from k_{p1} to k_{pn}) were re-paired to be the seeding pairs of the third round to generate the final keyword predictions. In theory this process could be repeated until no unique predictions remained, but in practice we found three rounds were sufficient for most domains. In terms of search trees, the breadth is determined by the number of seeding words and

the depth by the number of rounds of snowballing. If there are (n) unique predictions from the second round, then the seeding word pairing possibility would be n(n-1)/2, according to the previous formulas.

$$\mathbf{S}_{(\mathbf{k}_{p1},\mathbf{k}_{p2})} = \boldsymbol{f}_{GS}(\mathbf{k}_{p1},\mathbf{k}_{p2}) = \left\{\mathbf{k}_1^{p1,p2},\mathbf{k}_2^{p1,p2},\ldots,\mathbf{k}_{(n_{p1,p2}-1)}^{p1,p2},\mathbf{k}_{n_{p1,p2}}^{p1,p2}\right\}$$
$$\vdots$$
$$\mathbf{S}_{(\mathbf{k}_{p(n-1)},\mathbf{k}_{pn})} = \boldsymbol{f}_{GS}(\mathbf{k}_{p(n-1)},\mathbf{k}_{pn}) = \left\{\mathbf{k}_1^{p(n-1),pn},\mathbf{k}_2^{p(n-1),pn},\ldots,\mathbf{k}_{(n_{p(n-1),pn}-1)}^{p(n-1),pn},\mathbf{k}_{n_{p(n-1),pn}}^{p(n-1),pn}\right\}$$

4. Results

This automated methodology for generating rich ontology was applied against engineering sector application (WMCCM Collaborative marketplace), and an analysis of the resulting network was conducted.

4. 1. Primary Data

Three pairs of initial seeding words (drilling & cutting, milling & sawing, and turning & grinding) to represent the "machining" domain were obtained from the WMCCM project team. From these, 10,660 unique terms with 266,176 relationships among them were automatically generated using the procedure described in section 3. Previously WMCCM had used traditional manual processes to collect 862 unique concepts with 2,126 relationships from both SIC and domain experts. The new ontology contained fifty times more terms, and more than a hundred times the number of internal relationships compared with the original WMCCM ontology.

These terms and their relationships formed a "concept" network of terms. This network is similar to many social networks and there are well established social network analysis methods which can be applied to the collected data to conduct ontological analysis.

4. 2. Ontological Analysis

Ontological analysis enabled:

- Finding the "roots" – representatives of the network;
- Clarifying links between new domain terms and the "roots";
- Clustering of sub-trees and their defining boundaries and of the whole network.

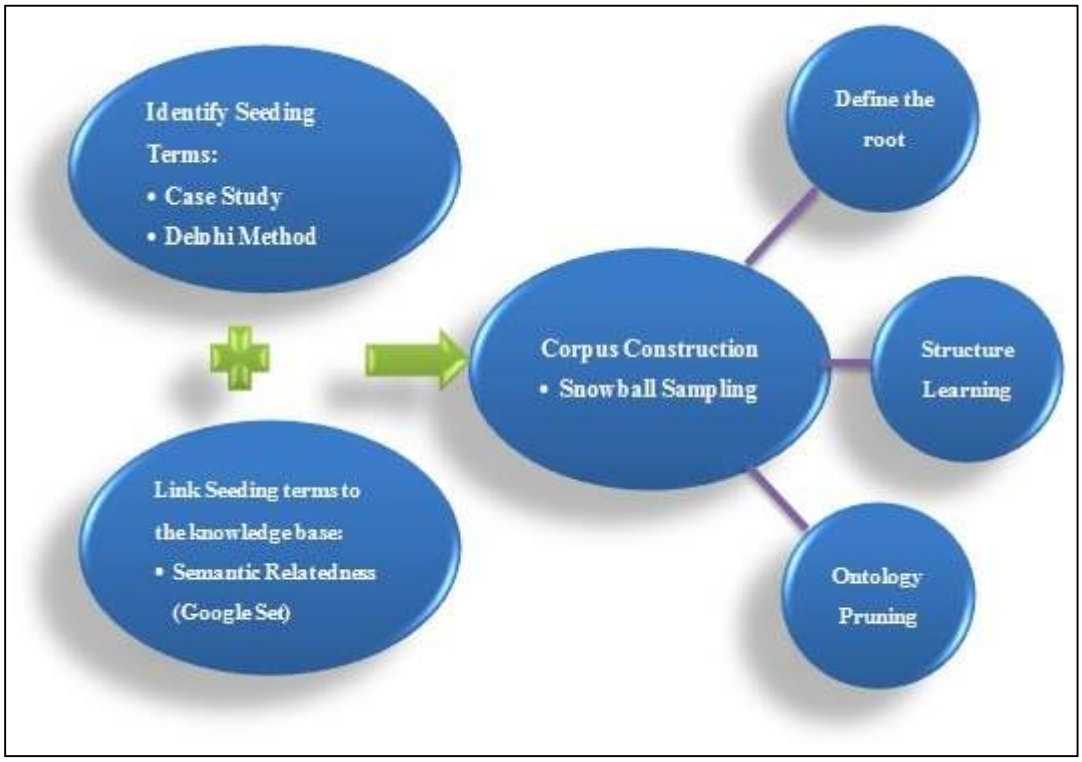

Figure 6: Detailed Techniques for Linking Seeding Words to the Knowledge Base

4.2.1. Centrality Analysis

There were 10660 unique keywords in the prediction sets, and their occurrences varied from once to 3432 times. Those members who had been "derived" (linked by others) more times could be regarded as more representative of the group, or more "centrally" located within a concept. Such centralised terms are the super connectors among groups of keywords (analogous to key social network members) within the overall network[16].

The corpus construction described in the experiment resulted in **n(n-1)/2** sets of collections. To examine the centrality of a target member (*m*) in such a data structure, the calculation had to go through every collection to count the possible relations it has with all the possible seeding words. Thus, the centrality algorithm had two steps:

Firstly, verifying the existence of (*m*) in every collection or Set (*S*), under the conditions that Set (*S*) was not seeded by a pair of words including (*m*) itself. The existence of (*m*) in Set (*S*) was configured as $f_E(m,S)$ to generate a numeric value.

$$f_E(m,S) = \left\{ \begin{matrix} 1, m \in S \\ 0, m \notin S \end{matrix} \mid f_{GS}(m,k) \neq S \right\}$$
$$Where: \quad S = \forall \; S_{(k_{pi},k_{pj})} \mid 1 \leq i < j \leq n$$
$$And \quad m \in \{k_{p1}, k_{p2}, \ldots, k_{pn}\},$$
$$k \in \{k_{p1}, k_{p2}, \ldots, k_{pn}\}, and \; m \neq k$$

Then, the total connections of (*m*) in these sets are the aggregation of $f_E(m,S)$. This can be calculated as the centrality:

$$f_{Cn}(m) = \sum_{i,j} f_E(m, S_{(k_{pi}, k_{pj})}) \mid 1 \le i < j \le n$$

Among 10660 generated keywords, 3920 keywords only appeared once. A one-time appearance implies that the predicted word does not have close connections with the other keywords but remotely connects with only one pair. For the purposes of this research, we defined "one time appearance" as noise in the experiment. The remaining keywords are distributed as shown below:

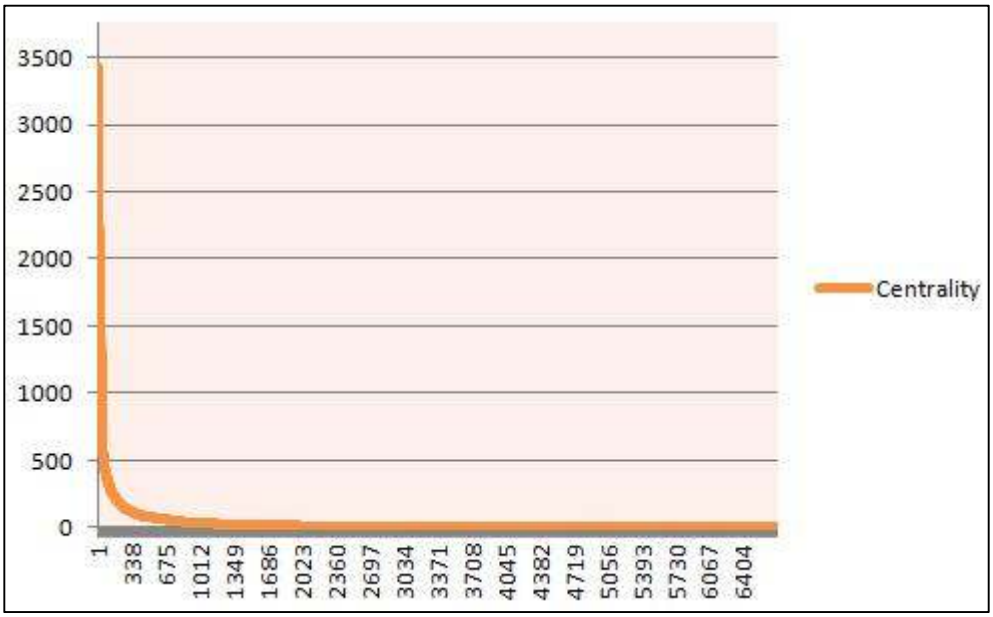

Figure 7: Keywords Centrality Analysis

This distribution is similar to a Poisson distribution. To understand more about the curve, we could cut it into 3 pieces by tangent (**y = -x)**. Then the curve would be divided into three distinguished zones (Figure 8):

1. Curve 1 (definition zone) presents a fully connected top zone with highly centralised members. Mathematically, these keywords appeared much more often than the other members outside the zone
2. Curve 2 (description zone) shows a fast drop that indicates those keywords used quite often as descriptors in the domain. Their centralities were lower than the top definition zone, but most of them were connected to top zone members.
3. Curve 3 (connection zone) includes those low centralised keywords mentioned around the concept, but not necessarily a part of the concept, although they do have some connection with the some of the words in the definition or description zone.

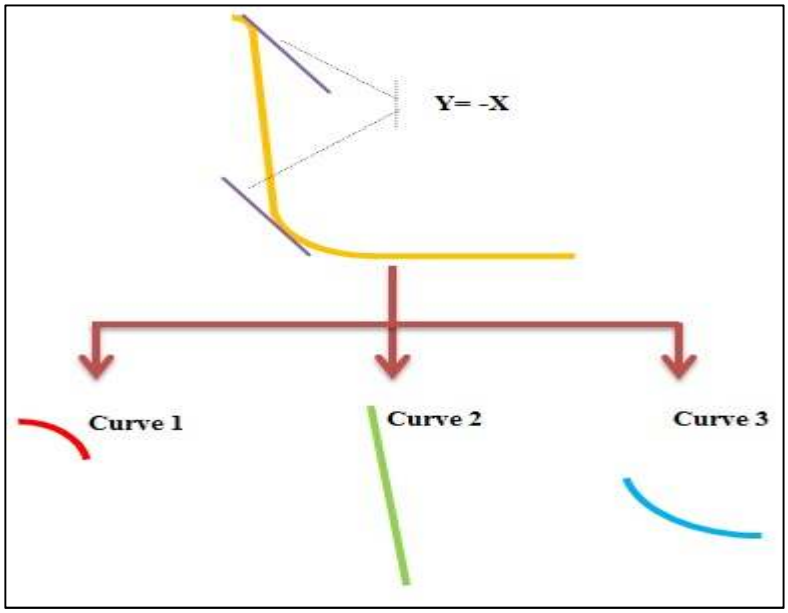

Figure 8: Cut-off Points

4.2.2. Closeness Analysis

"Closeness" analysis takes concepts within a domain as observation objects to measure how close concepts are to each other. Unlike centrality analysis, it counts the connections to a concept from another concept. Closeness could be treated as the relevant connective power between concepts. This relevant power can indicate the "closeness" between concepts. In addition, the sum of connections provided a numeric value, and it could be converted (a simple method is to use reciprocal) to a value from 0-1, which could represent the distance between conceptual clusters.

In this research, the closeness investigated how important a seeding word (*k*) was in predicting (*m*), and in semantic relatedness terms, how much did seeding word (*k*) determine the appearance of prediction (*m*) in the domain. Centrality analysis defined $f_{Cn}(m)$ to track (*m*) appearances in all the prediction sets, regardless of their seeding words. If seeding words were considered, for example a seeding word *k*, $f_{cl}(m,k)$ can calculate *m*'s appearances via a traversal of these sets, based on *k*.

$$f_{cl}(m,k) = \sum_{i=1}^{n} f_E(m, S_{(k_{pi},k)})$$

Then, the decisive power of seeding word *k* on predictions *m* could be presented as a closeness distance $f_d(m,k)$. The greater $f_d(m,k)$ is, the greater the decisive power *k* has to predict *m*.

$$f_d(m,k) = \frac{f_{cl}(m,k)}{f_{Cn}(m)}$$

The result of practical closeness analysis on the corpus confirmed that different seeding words had different decisive powers over the number of appearances of a target word. A quantified value helped to refine the zone definition from centrality analysis, as centrality analysis can only conduct zone specification from a structure perspective.
The new methodology generates connections between different terms that are weight specified directional relationships (like vectors) based on the "closeness" value. Such relationship expresses the binary relationship more richly than simple weightless connection. For example, Table 1 demonstrates the relationship between several terms to the concept "turning".

Seeding Words (k)	Predict(m)	$f_{cl}(m,k)$	$f_{Cn}(m)$	f_d(m,k)	Relevant Distance
Reaming	Turning	115	2664	0.043168	1
Tapping	Turning	106	2664	0.039790	1.084906
Threading	Turning	97	2664	0.036411	1.185567
Conventional turning	Turning	93	2664	0.034910	1.236559
Screw cutting	Turning	93	2664	0.034910	1.236559
Drilling	Turning	79	2664	0.029655	1.455696
Centering	Turning	79	2664	0.029655	1.455696
Micro drilling	Turning	72	2664	0.027027	1.597222
Deburring	Turning	67	2664	0.025150	1.716418
Cutting	Turning	65	2664	0.024399	1.769231
CNC Machining	Turning	26	2664	0.009760	4.423077
Thread rolling	Turning	22	2664	0.008258	5.227273

Table 1: Weight Specified Relationship

Drilling and centring can be associated with either turning or milling. The "distracted" linkage towards both turning and milling may reduce the strength of the relationships towards either of them. Therefore, they appeared to be less strongly related to turning process.

4.2.3. Betweenness Analysis

"Betweenness analysis" identifies those members whose importance may be missed by centrality and closeness analysis but who bridge the gaps between concept clusters. Betweenness analysis finds those individuals or groups who have concurrent membership in overlapping concepts, so the relations between concepts become clearer. In this research, members with significant "Betweenness" factors were found via the following method:

1. Reference to the closeness addressed those members with a low closeness in the network; this meant that such concept clusters were semantically further than

others. In this research, special attention was paid to those numbers that are remotely positioned in both directions. For instance, the traversal of f_d could address predictions m_1 and m_2, where:

$$f_d(m_1, m_2) \to 0 \quad \text{and} \quad f_d(m_2, m_1) \to 0$$

Addressing this sort of relationship was the key to clarifying the conceptual clusters, especially when both m1 and m2 were highly centralised members. It provided numerical figures to draw boundary between m1 and m2.

2. But there may exist a prediction k which is decisive for both m1 and m2:

$$f_d(k, m_1) \to \max_{1 \le i \le n} f_d(k, k_{pi})$$
$$\text{and} \quad f_d(k, m_2) \to \max_{1 \le i \le n} f_d(k, k_{pi})$$

Such *k* connected m_1 and m_2 from *k*'s view point. The existence of such keyword shows that bridging concepts exists and could be located. It also indicates that the peripheral players of a network should not be omitted, since they may be the bridge to other networks.

The analysis revealed that this method of analysis was able to create well positioned "betweenness" measures between members. For example, table 2 shows that "folding" and "honing" in the generated engineering ontology are not particularly close to each other. However, there was a member "tool grinding" which is tightly connected to both of them.

Seeding Words (k)	Predict (m)	$f_{cl}(m,k)$	$f_{Cn}(m)$	$f_d(m,k)$
Folding	Honing	3	2121	0.001414
Honing	Folding	1	1131	0.000884
Tool grinding	Honing	83	2121	0.039132
Tool grinding	Folding	58	1131	0.051282

Table 2: Example of the Betweenness Analysis in the Engineering Ontology

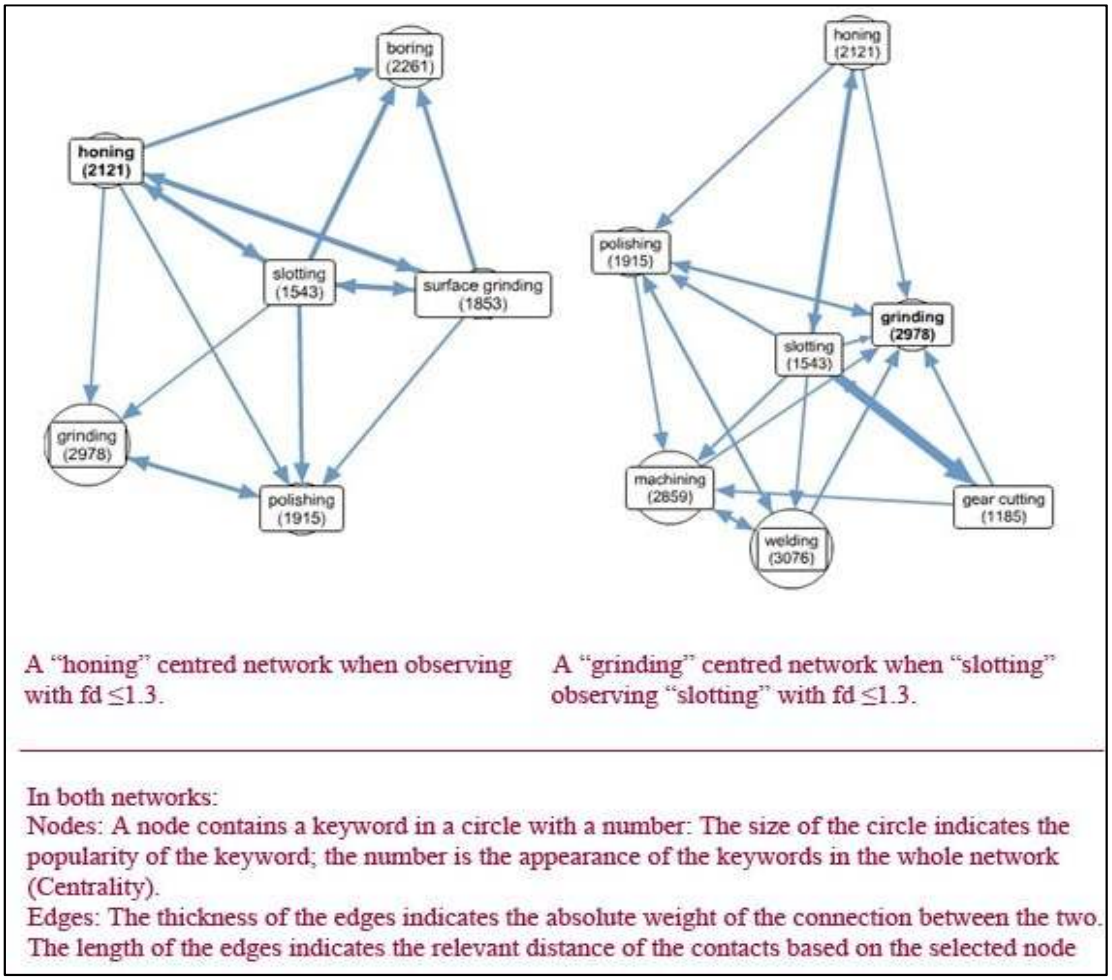

Figure 9: Illustration of the Engineering Ontology Network

5. Discussion

The research also investigates that the process is repeatable, that cut-off points were set reasonably, that the final output serve our research objectives, and that the research could be applied to real life environment.

5.1. Zones Explanation

5.1.1. Connection Zone

The ground level connection zone contains "long tail" terms nominated by the terms in the two upper levels. Terms in the ground level did not necessarily describe the main concepts accurately, but they were connected to the concepts or concepts descriptions to some extent in the domain context. For example, "food processing" was identified as a connection zone member in the new engineering ontology. Practically, such a connection zone member does have a relationship with the main concepts. However, the frequency of appearance of the terms in this zone was the lowest in the three zones. These third zone terms were valuable from other perspectives: in terms of structural clarification such members could be boundary players

and from a cross domain viewpoint, they may be the brokers from the target domains to related domains.

5.1.2. Description Zone

This zone is populated by popular concepts or terms defining in more detail the concepts from the top zone. Observation of these concepts or terms revealed that many of them were phrases containing concepts or their thesaurus from the top definition zone. At this level, terms were inevitably connected to the relevant concepts at the top level but were not as important as them (lower centrality value). For example, “drilling” is a core concept in the new engineering ontology; its directly linked concepts “gun drilling” and “cross drilling” are description zone members.

Members in the description zone have at least one direct connection to a few but not all of the top zone members, and additionally they have limited connections with each other. Not being able to form a complete network is a distinguishing characteristic of the remainder of the network members. Incomplete network also implies separation of their corresponding concepts (or conceptual clusters), thus borders could be drawn based on such disconnectivity. Although not fully connected, these members can reach all top level members and most of the other descriptive members within three steps as required by network reach analysis.

5.1.3. Definition Zone

Compared with other zones, the keywords in definition zone appear more often, and they are thus the keywords that define the domain(s) most explicitly.

In the definition zone, members cover most of WMCCM categories and the UK SIC codes for the engineering area. For example, [5] describes machining (first column in Table 3) as: This class includes:

- cutting, boring, turning, milling, eroding, planing, lapping, broaching, levelling, sawing, grinding, sharpening, polishing, welding, splicing etc. of metalwork pieces
- Cutting of and writing on metals by means of laser beams.

Keyword	Centrality	Keyword	Centrality
Drilling	3432	Centering	1862
Welding	3330	Conventiona l turning	1852
Milling	3157	Slotting	1776
Machining	3148	Electroformin g	1747
Grinding	3128	Screw cutting	1741
Cutting	3012	Tool grinding	1667
Tapping	2879	Gear shaping	1660
Sawing	2824	Stamping	1644
Turning	2789	Micro drilling	1643
Painting	2771	Finishing	1511
Assembly	2765	Fabrication	1490
Punching	2685	Gear cutting	1482
Bending	2468	CNC Machining	1456
Boring	2408	Rolling	1263
Deburring	2344	Heat treating	1216
Forming	2331	Laser cutting	1206
Honing	2305	Folding	1169
Broaching	2270	Plating	1106
Shearing	2192	Notching	1095
Polishing	2144	Custom fabrication	1002
Threading	2125	Engineering	919
Reaming	2080	Powder coating	912
Surface grinding	2077	Design	912
Cylindrical grinding	1919	Thread rolling	901
Surfacing	1896	Plasma cutting	856

Table 3: Definition Zone Members

Nine out of fifteen keywords in the SIC definition are covered by the definition zone, with the remainder covered by the lower zones (4 by the description zone and 2 by the connection zone). In addition, the research generates all the WMCCM categories that exist in the set. WMCCM proposed 22 concepts in the definition zone (second column in Table 4).

With the new ontology, 16 out of 22 of these concepts were covered by the definition zone and another three have high centrality in the description zone, with the rest covered by the connection zone. Moreover, the prediction set generated covers more domain space than both the SIC and WMCCM ontology. The results provide evidence that they are not only accurate, but also have a wider coverage than the standard code (see Table 4).

SIC	WMCCM Ontology	New Ontology	Centrality
Boring	Boring	Boring	2408
Broaching	Broaching	Broaching	2270
	CNC Laser Cutting	Laser Cutting	1206
	CNC Machining	CNC Machining	1456
	CNC Milling	CNC Milling	511
	CNC Turning	CNC Turning	405
Cutting	Cutting	Cutting	3012
	Drilling	Drilling	3432
Eroding		Eroding	64
	Fettling	Fettling	2
	Gear Cutting	Gear Cutting	1482
Grinding	Grinding	Grinding	3128
	Hobbing	Hobbing	2305
	Manual Machining	Machining	3148
Lapping		Lapping	289
Levelling		Levelling	25
Milling	Milling	Milling	3157
Planning		Planning	58
Polishing		Polishing	2144
	Profiling	Profiling	143
Sawing	Sawing	Sawing	2824
	Splining	Splining	37
Sharpening		Sharpening	92
Splicing		Splicing	2
	Tapping	Tapping	2879
	Thread Grinding	Thread Grinding	42
	Threading	Threading	2125
Turning	Turning	Turning	2789
Welding	Welding	Welding	3330

Definition Zone Description Zone Connection Zone

Table 4: Ontology Content Comparison

5.2. Repeatability

The similar experiment has also been conducted for the other domains to assess if the appearance curve will remain the same shape. This showed the same trend as engineering: a fairly short definition zone, a sharp drop description zone and a very long tail connection zone. Such repetition of the curves indicated that the predictions do maintain the same trend and the experiment is repeatable.

5.3. Fault Tolerance

Another valuable contribution of the research is that it has some fault tolerant ability. Originally, the research was designed to have three pairs of keywords to avoid potential misdirection by a badly chosen term. Three pairs will allow one pair to be misleading, but will still have 66.7% outputs towards to the right direction in theory.

In fact, we did have a bad sample in our experiment: one of our original chosen words was “hobbing”, and its appearance was only 120, which made it fall into the connection zone. But contrarily, this expresses the fault tolerance ability of the system: ‘hobbing’ is recognised in connection zone, so it has quite limited affection to the other 2 more important zones.

5.4. Optimisation of Current Process

The derived ontology for this research was built to solve practical problems in information categorisation for WMCCM. Monitoring mechanism was implemented to compare the performance of the original engineering ontology used by WMCCM and the ontology developed through this research. More than 5000 engineering tenders were processed through the system every day. Figure 9 demonstrates that the categorisation system has been improved by adopting the new ontology:

- The new ontology filter was triggered by more than 91% of the input information, compared to 82% triggered the existing WMCCM ontology.
- Among those filtered items, 77% of the information had appropriate categorisation by the new ontology, compared to only 51% were correctly categorised by the existing one, which was due to insufficient internal relationships within the existing ontology.

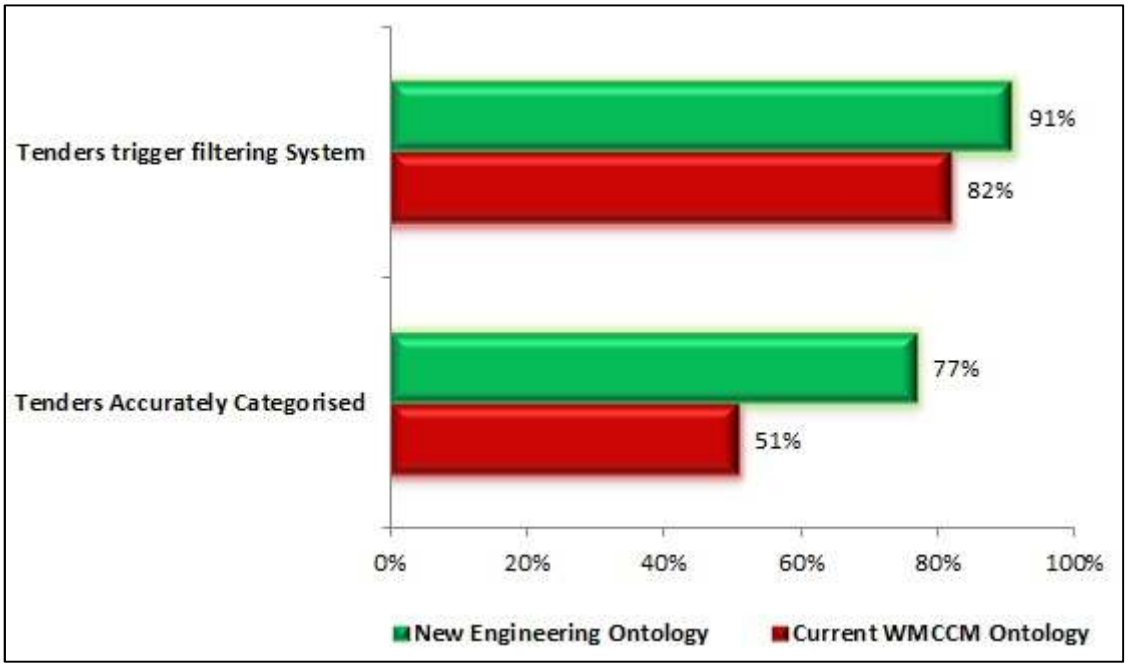

Figure 10: Practical Evaluation of the New Engineering Ontology

Practical evaluation proved that the new derived ontology can be fitted to the desired automated system and provided better categorisation results. More importantly, the new ontology could be fitted to an existing fixed ontology by adding the generated rich concepts and relationships as conceptual descriptions (Such descriptions only supplement additional terms and relationships without changing the ontological structure).

6. Conclusion

Good ontologies can play a key role in self-help systems for "intelligent" processing and categorisation. Through the investigation of WMCCM ontology and other relevant ontologies in the engineering and manufacturing domain, the need was identified to quickly, reliably and economically generate ontologies that are able to provide the breadth and depth of coverage required for the given domain.

A new ontology development methodology has been proposed to address those needs, and the derived ontology has been implemented and evaluated to improve the current ICT system's categorisation. The derived ontology addresses the issues regarding the cost of generating ontologies with sufficient scope and relationships richness. It has been demonstrated that a rich multi-disciplinary ontology can be built with only three pairs of seeding words provided by a domain expert using semantic-relatedness-based tool. This ontology has a high breadth and depth of concept coverage and derives internal relationships to form a network structure. The evaluation of the derived ontology has demonstrated that it has performed better in the automated information categorisation applications than the industry code and the current ontology adopted by WMCCM.

7. References

[1] G. van Heijst, A.T. Schreiber, B.J. Wielinga, Using explicit ontologies in KBS development, International Journal of Human-Computer Studies, 45 (1997) 183-292.
[2] R. Mizoguchi, J. Van Welkenhuysen, M. Ikeda, Task Ontology for Reuse of Problem Solving Knowledge, in: N.J.I. Mars (Ed.), Towards Very Large Knowledge Bases, IOS Press, Amsterdam, 1995, pp. 60-72.
[3] M. Swift, N. Armoutis, J. Bal, M. Molfetas, The Formation of Virtual Organisations to Address Complex Tenders through a Collaborative Commerce Marketplace, in, Proceedings of the 13th International Conference on Concurrent Enterprising, Sophia-Antipolis, France, 2007.
[4] A.M. Fairchild, B. De Vuyst, Coding Standards Benefiting Product and Service Information in E-Commerce, in, Proceedings of the 35th Hawaii International Conference on System Sciences, IEEE Computer Society, Hawaii, 2002.
[5] L. Prosser, UK Standard Industrial Classification of Economic Activities (SIC 2007), in, 2007.
[6] O. Corcho, A. Gómez-Pérez, Solving Integration Problems of Ecommerce Standards and Initiatives through Ontological Mappings, in: A. Gómez-Pérez, M. Grüninger, H. Stuckenschmidt, M. Uschold (Eds.), IJCAI'01 Workshop on Ontologies and Information Sharing, CEUR-WS.org, Seattle, Washington, 2001, pp. 131-140.
[7] E.K. Jacob, Classification and Categorization: A Difference that Makes a Difference, LIBRARY TRENDS, 52 (3) (2004) 515–540.
[8] E. Mayr, The growth of biological thought: diversity, evolution, and inheritance, The Belknap Press of Harvard University Press, Cambridge, Massachusettes, 1982.
[9] EAGLES, Expert Advisory Group on Language Engineering Standards. Preliminary Recommendations on Semantic Encoding, in, 1998.
[10] M.d. Kunder, Size of The World Wide Web, in, 2012.
[11] A. Gulli, A. Signorini, The indexable web is more than 11.5 billion pages, in, Special interest tracks and posters of the 14th international conference on World Wide Web, ACM, Chiba, Japan, 2005, pp. 902-903.
[12] S. Tong, J. Dean, System and methods for automatically creating lists, in, USA, 2008.
[13] H.A. Linstone, M. Turoff, The Delphi Method: Techniques and Applications, Addison-Wesley, Reading, Mass., 1975.
[14] M.J. Salganik, D.D. Heckathorn, Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling, Sociological Methodology, 34 (2004) 193-239.
[15] O. Frank, Network Sampling and Model Fitting, Cambridge University Press, New York, 2005.
[16] L. Katz, A new status index derived from sociometric analysis, Psychometrika, 18 (1) (1953) 39-43.