# A Thesaurus-Guided Framework for Visualization of Unstructured Manufacturing Capability Data

Farhad Ameri, William Bernstein

# A Thesaurus-guided Framework for Visualization of Unstructured Manufacturing Capability Data

Farhad Ameri[1], William Bernstein[2]

[1]Engineering Informatics Lab, Texas State University, San Marcos, U.S.A
ameri@txstate.edu
[2]Systems Integration Division, National Institute of Standards and Technology (NIST) Gaithersburg, U.S.A
william.bernstein@nist.gov

**Abstract.** Manufacturing companies advertise their manufacturing capabilities and services on their company website using unstructured natural language text. The unstructured capability data published on the web is a rich source of formal and informal manufacturing terms and knowledge patterns. Through systematic mining of a large collection of capability text, new semantic models and knowledge graphs can be extracted that can be used as the stepping stone of more formal ontologies. The objective of this research is to develop a framework for better understanding, analyzing, and summarizing manufacturing capability data that is available on the websites of manufacturing companies. The findings can support supply chain decisions and may result in the discovery of new trends and associativity patterns in the data. The focus of this paper is on demonstrating how visual analytics (VA) tools can be used for gaining insights into manufacturing capability and the associativity pattern among the capability entities labeled by various terms. A visual analytics system, named Jigsaw, is used for exploring the connections between various entities such as process, material, industry, and equipment across the documents in the experimental dataset.

**Keywords:** manufacturing capability, thesaurus, visual analytics, Jigsaw

## 1 Introduction

Manufacturing capability can be defined as the "firms' internal and external organizational skills, resources, and functional competencies to meet the requirement of the changing economic environment" [1]. Manufacturing capability is a multi-faceted entity represented through different criteria such as quality, processing capability, production capacity, flexibility, product innovation capacity, and performance history. Capability data is often presented in both structured and unstructured formats. The structured capability data can be found in proprietary application databases connected to the enterprise legacy systems of the company. The unstructured capability data is available in plain text form, often written in a loose narrative style, on the websites of manufacturing companies. The structured data is more amenable to complex query and quantitative analysis whereas, the unstructured data cannot be readily analyzed and understood especially when dealing with very large quantities of capability text. The objective of this research is to develop a framework for better understanding, analyzing, and summarizing the manufacturing capability data that is published on the websites of manufacturing companies. The findings can support supply chain decisions and may result in discovery of new trends and associativity patterns in the data. For example, it would be useful to classify and cluster manufacturers based on their similarities with respect to process and material capabilities and competencies proclaimed on their website. This would provide supply chain decision makers with better insight into the capabilities of prospective suppliers during supplier selection and evaluation process. Another motivation for analyzing unstructured capability data is that there is a wealth of manufacturing knowledge hidden in the text which, if mined systematically, can result in discovering interesting rules and patterns that were otherwise unknown. Such trends and patterns could inform early upstream decisions in the product design lifecycle.

In previous research, a supplier classification method based on Naïve Bayes text classification technique was implemented. It was observed that capability-based classification can result in accurate supplier groups with unique capability attributes [2]. This paper focuses on applying visual analytics to large collections of capability data. Visual Analytics (VA) combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets [3] . VA has been used in the manufacturing domain for visualizing different types of manufacturing data such as simulation, quality control, and distribution and sales data [4-6]. The literature search, however, did not reveal any work related to applying visual analytics for exploration of

manufacturing capability data. There are multiple VA-based tools and systems available to conduct visual exploration of domain-specific data [7]. In this research, Jigsaw [8] is used as the VA-based tool due to its powerful document analytics features and functions and also its simple, intuitive interfaces. The primary unit of analysis in Jigsaw is an "entity". In manufacturing capability text, examples of entities include, processes, materials, industries, and equipment. One of the novelties of the proposed text analytics framework is that it is guided by a formal thesaurus of manufacturing capability terms that provides the dictionary of entities, or concepts, required as input by Jigsaw. Additionally, an automated Entity Extractor Tool (EET) is developed in this work to streamline the data collection and preparation process.

This paper is organized as follows. Section 2 provides a brief description of the Manufacturing Capability Thesaurus. The data collection and preparation method is described in Section 3. The visual analytics examples and results are presented in Section 4. The paper ends with the conclusions in Section 5.

## 2 Manufacturing Capability Thesaurus

In the proposed visualization method, the unit of analysis is an entity or a concept. A concept is "a unit of thought, formed by mentally combining some or all of the characteristics of a concrete or abstract, real or imaginary object. Concepts exist in the mind as abstract entities independent of terms used to represent them" [9]. For example, the concept of operating a machine tool manually can be labeled by terms such as *manual machining*, *conventional machining*, or *traditional machining*. There are thousands of concepts in the manufacturing domain that provide domain experts with the necessary vocabulary when describing the manufacturing capability of a machine, a production plant, or a supply chain composed of multiple companies. Most text analytics techniques require a predefined list or dictionary of concepts that can be used for tokenizing the documents in a given data set. The dictionary of concepts is sometimes generated automatically using machine learning techniques based on a corpus. However, the caveat being that automatically-generated dictionaries are often contaminated by noise. Furthermore, creation of a dependable corpus is a demanding task. In this work, a semi-automated method is used for creation of a high quality thesaurus of capability concepts, called the Manufacturing Capability (MC) Thesaurus.

The MC Thesaurus contains the typical concepts, and their associated terms, that are often used for describing the manufacturing capabilities of suppliers. Contract manufacturers may use terms and phrases such as *precision machining*, *tool and die making*, *turnkey services*, *build-to-order manufacturing, small to large volume production runs* to explicitly describe their technical capabilities, capacities, expertise, and/or services. Also, they may provide examples of parts they have produced or industries and customers they have served in the past to advertise specific abilities or skills. Another complicating factor is the short hand vocabulary in every industry that is only meaningful to the individuals that are immersed in that industry. A thesaurus can be used for semantically linking the seemingly disparate terms in different subcategories of manufacturing industry, thus reducing terminological ambiguities. In the presence of a comprehensive thesaurus of manufacturing capability terms, it is possible to readily translate each website into a concept vector model that is more suitable for quantitative analysis. Also, the MC Thesaurus provides the necessary entity dictionaries that are required as input to different document analytics tools such as Jigsaw. The MC thesaurus is a formal thesaurus in a sense that it uses SKOS [10] (Simple Knowledge Organization System) for syntax and semantics. As an open-source thesaurus with explicit semantics, the MC thesaurus can be shared as linked data to be reused by dispersed users.

Each concept in SKOS has exactly one preferred label (*skos:prefLabel*) and can have multiple alternative labels (*skos:altLabel*). For example, as can be seen in **Fig. 1**, *Screw Machining* is the alternative label for *Swiss Machining* as it is used frequently for referring to the same concept.

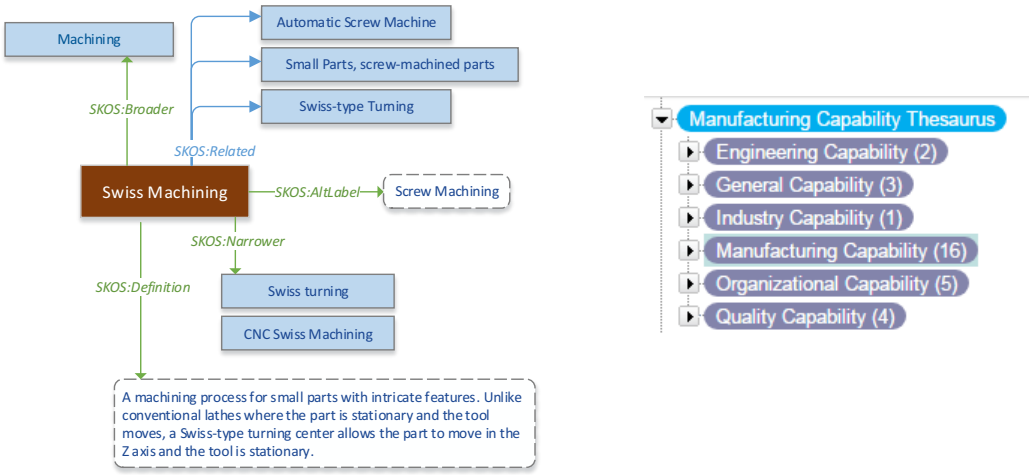


**Fig. 1**: Concept diagram of Swiss machining process in MC thesaurus (left), the concept schemas of MC thesaurus (right)

MC thesaurus currently contains more than 800 capability concepts categorized under six concept schemes as listed in Fig. 1(right). Also, concepts can be made related to one another using *skos:related* relation. For example, *Swiss Machining* is related to *Small Parts* and *Automatic Screw Machine*. By connecting the concepts using *skos:related* relation , the MC thesaurus becomes a network of semantically related concepts.

## 3 Data Collection and Preparation

The raw data that is used in this research is *Unstructured Capability Data*. For this research, *Unstructured Capability Data* is defined as a text, in natural language, that provides direct or indirect pointers to the technical and non-technical capabilities of manufacturing companies. In this research, the capability data is collected from the websites of North American companies in contract manufacturing industry. The core services provided by the target suppliers are primarily in the area of precision CNC machining. The concept vector for each capability text is generated using the *Entity Extractor Tool* (EET) that is developed for this research. The entities are essentially the concept labels from the MC thesaurus that occur in the text. The EET receives the plain text, that contains capability data, or the URL of a supplier website as the input, and generates the list of detected concepts and their frequencies as a CSV file. Fig. 2 shows the screenshot of the input page of the Entity Extractor Tool. If a URL is provided by the user as the input, the EET creates a preview list of the pages from the same domain as the provided URL depending on the user-specified crawling depth. Usually a crawling depth of one page returns most of the important sub-pages that are directly hyperlinked to the homepage but occasionally a depth of 2 or more might be required to capture the useful text. The user can then eliminate the pages that seem irrelevant to capability information by simply unchecking the box in front of them. Typically, the pages under Capabilities, Services, or Processes link contain useful data with capability relevance that can be included in the entity extraction and analysis processes.

Fig. 2 (right) shows the list of pages returned by the crawler for the example company. Note that, in this case, the user has checked only the pages that most likely contain capability data. The text for all crawled pages is gathered and aggregated as a single *document*, with a ".txt" extension, and then it is submitted for analysis. The analysis entails identifying MC Thesaurus concepts and their frequencies in the document. Fig. **3** shows the result of analysis for the example company. As can be seen in this figure, CNC Machining Center, Aerospace Industry, and Defense are the concepts with high frequency for the example supplier. Therefore, it can be concluded that, most likely, this company provides CNC machining services for aerospace and defense industries.

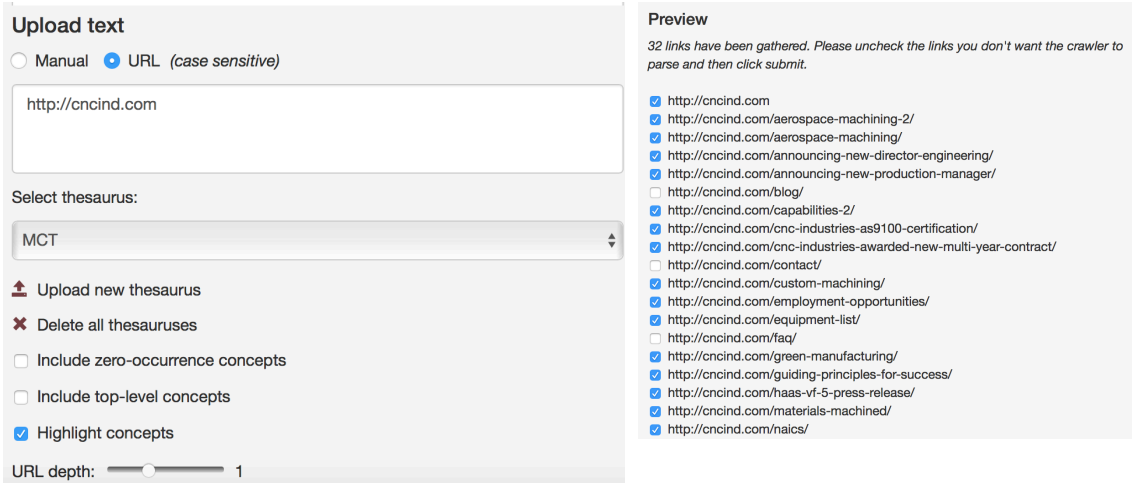


**Fig. 2:** The input screen of the EET (left), The list of pages retuned for a given supplier. Only the relevant pages are selected to be included in the entity extraction process (right.)

The entity extractor tool is capable of identifying both preferred and alternative labels for the MC concepts. The terms highlighted in green are the preferred labels and the terms highlighted in red are the alternative labels. The frequencies of alternative labels are added to the frequency of their corresponding preferred label. The user can export the vector model of each supplier as a CSV file with two columns, namely, concept (preferred label) and its frequency (number of occurrence in the extracted text).

Once the vector models, or the facets, of the documents are created, they can be used for different types of capability analysis. This paper focuses on visual analysis of both the generated vector models and the raw text documents.

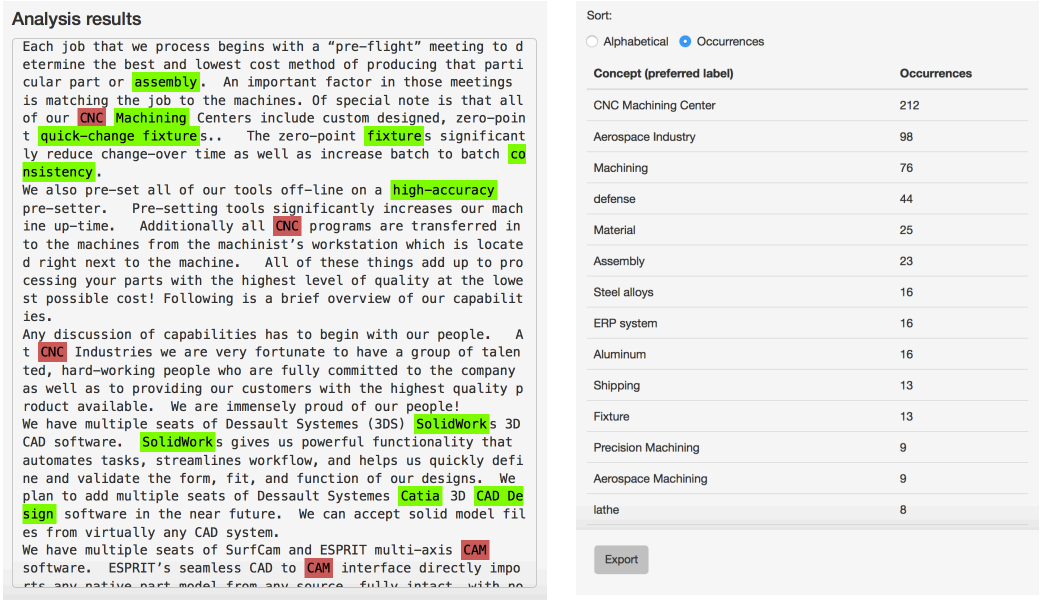


**Fig. 3:** The result of crawling together with the generated vector model

## 4 Visual Analytics

An experiment was designed and conducted to visualize capability data for a dataset collected from Thomas Net[1] website which is a web portal for manufacturing sourcing. For this experimentation, 40

[1] http://www.thomasnet.com

suppliers in contract machining industry were selected and their capabilities and qualifications were verified manually. The sample was intentionally formed such that half of the selected suppliers are qualified as heavy machining suppliers and the other half are qualified as complex machining suppliers.

## 4.1 Visualization of Concept Vector Models

The vector model of each text contains the manufacturing capability features of the supplier corresponding to the capability text. Enabled by the EET, it is possible to quickly generate a large collection of vector models (CSV files). The vector model can be directly used as the input to visualization tools. As an example, Fig. 4 shows a scatter diagram built based on a sample of 20 suppliers, randomly selected out of the starting pool of 40 suppliers. This diagram shows a combined view of heavy machining and complex machining capabilities. The capability score for each supplier is calculated based on the ratio of the number of the concepts pointing to heavy machining (or complex matching) capabilities for the supplier over the number of all heavy machining (or complex machining) concepts available in the MC thesaurus.

$$Score_i^{HM} = \frac{nc_i^{HM}.N_i^{HM}}{nc_T^{HM}}$$

Where, $Score_i^{HM}$ is the Heavy Machining (HM) score for the ith supplier, $nc_i^{HM}$ is number of HM concepts for the ith supplier, $nc_T^{HM}$ is the total number of HM concepts in the thesaurus, and $N_i^{HM}$ is a normalizing factor that takes into account the total number of concepts detected for the ith supplier.

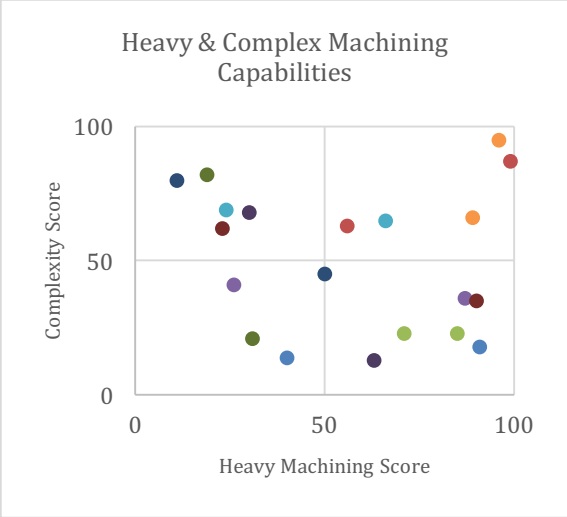


**Fig. 4:** Scatter diagram showing the combined view of complexity and heavy machining capabilities

## 4.2 Document visualization using Jigsaw

Although the vector model of capability documents provides an effective representation for computational analysis, it does not always provide a complete picture of supplier's capability. The completeness of the vector model directly depends on the completeness of the thesaurus. Therefore, it is necessary to supplement the vector-based analysis with more in-depth analysis that uses the entire document as the input. For this purpose, the Jigsaw analytics system, an existing tool developed at Georgia Tech, is leveraged in this section. Jigsaw is a visual analytics-based system that represents documents and their entities visually and displays the connections between entities across documents [8]. Entities are connected if they appear in one or more documents together. There are multiple mutually-coordinated visualization frames, called *Views*, that provide different perspectives onto data. List view, graph, view, scatter plot view, document cluster view, and grid views are among the most important views in Jigsaw. Jigsaw was originally developed with the objective of helping investigative analysts, in domains such as intelligence and law enforcement, build theories and formulate hypotheses based on a collection of documents.

In this work, Jigsaw is evaluated experimentally to assess its effectiveness when analyzing manufacturing capability documents. The same set of 40 suppliers used in the previous analysis was imported into a Jigsaw project as a batch of text files. Each capability text is treated as a document in this project. Four categories of entities, namely, *industry*, *material*, *process*, and *equipment*, were also created to build the underlying entity model of the project as shown in Fig. 5. Dictionaries of these entities were directly obtained from the MC thesaurus. This experiment was geared toward exploring connections between companies, industries, processes and materials. The numbers in parenthesis in the left image in Fig. 5 show the number of entities under each category that are observed in the documents. The entities that

are synonyms were aliased together. For example, Swiss Machining and Screw Machining were aliased together because they both point to the same concept in the MC Thesaurus.

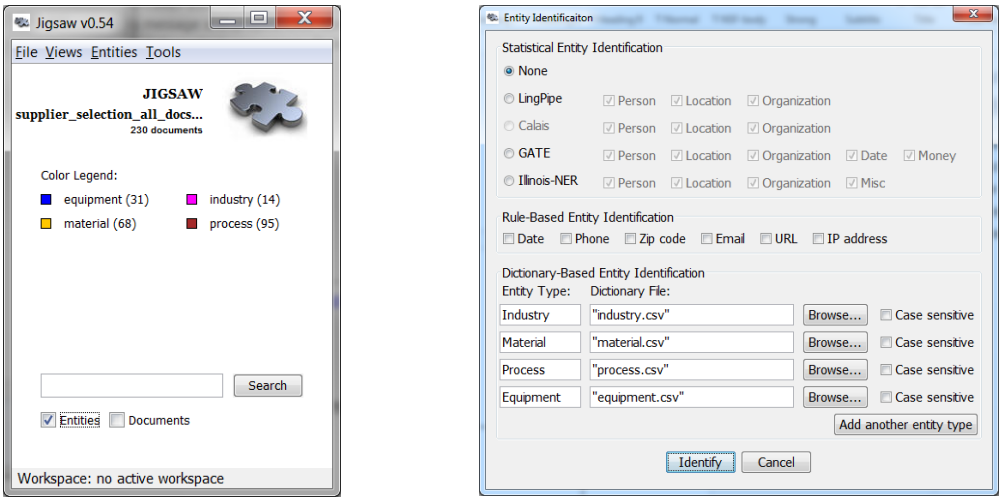


**Fig. 5:** Four categories of Entities were imported to Jigsaw based on the vocabulary extracted from MC thesaurus

Fig. 6 shows one possible list view built based on the input dataset and entity model. The list view provides multiple reorder-able lists of entities. Also, the documents themselves can be added as a list to visualize the links between the connections and entities. In Fig. 6, documents are listed in the third column from the left. The strength of connection between entities and documents are represented by different hue saturations of a common color. For example, in Fig. 6, the selected industry entity, aerospace, has a stronger connection with *machining* compared to *broaching* listed under the process list, hence has a more prominent single color hue saturation. The connection between aerospace industry and various engineering materials is also visualized in this view. For example, as expected, *aluminum* has a stronger connection with aerospace industry when compared to *high carbon steel*.

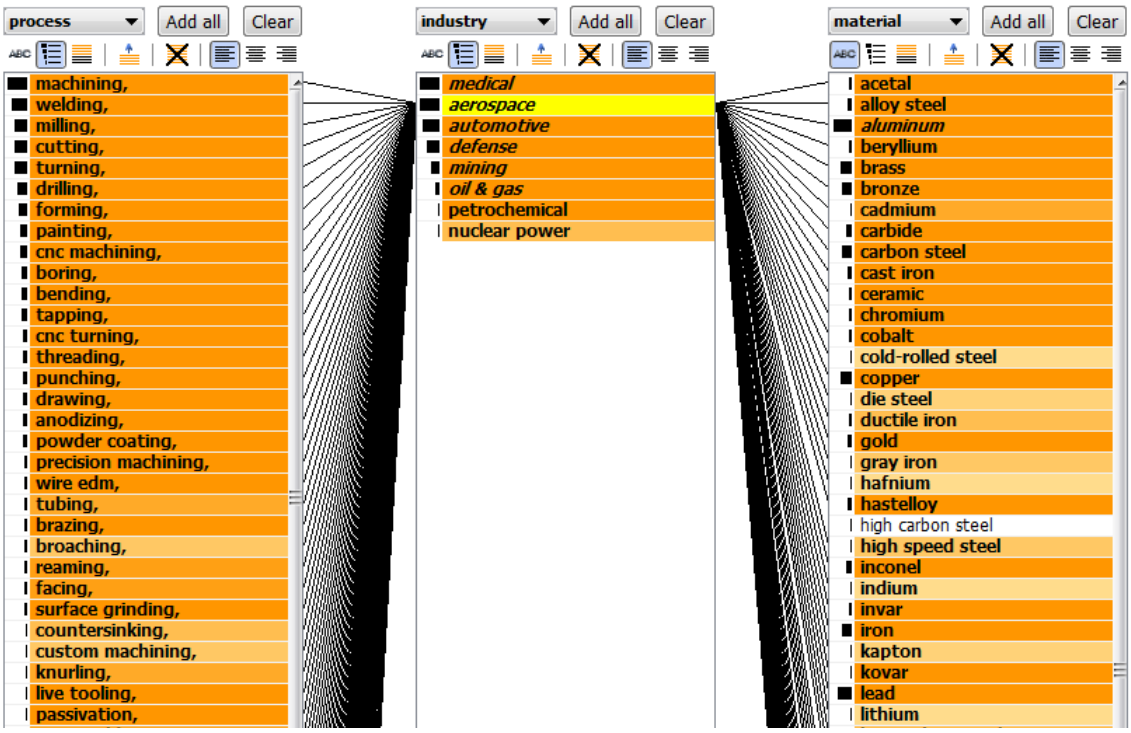


**Fig. 6:** List view in Jigsaw

Fig. 7 shows the graph view of the dataset. Each white rectangle in the graph view represents a document and the circle, colored nodes represent entities. An edge linking a document to an entity means that the entity is observed in that document. A mouse click on each document node expands or collapses the node to show or hide the links between the document and the entities. It should be noted that most of the edges in Fig. 7 are hidden for the purpose of demonstration and readability. The graph view can be rearranged to obtain more informative views. For example, in Fig. 8, the nodes are automatically arranged based on centrality to the document-type nodes. In this view, one can see that *automotive*, *tubing*, and *vertical machining center* are more central (i.e. relate to more documents) than other nodes of the same type, *nuclear power*, *open die forging*, and *automatic bar feeder*. This view presents an overview of all the relationships and help identify single entities that only relate to a single document, e.g. *open die forging*.

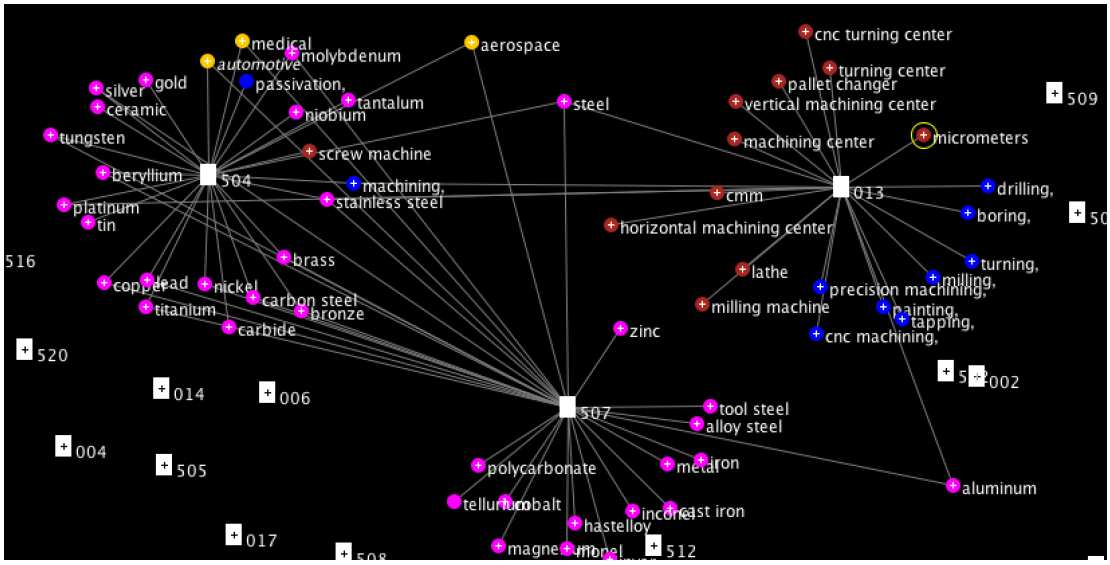


**Fig. 7:** Graph view in Jigsaw

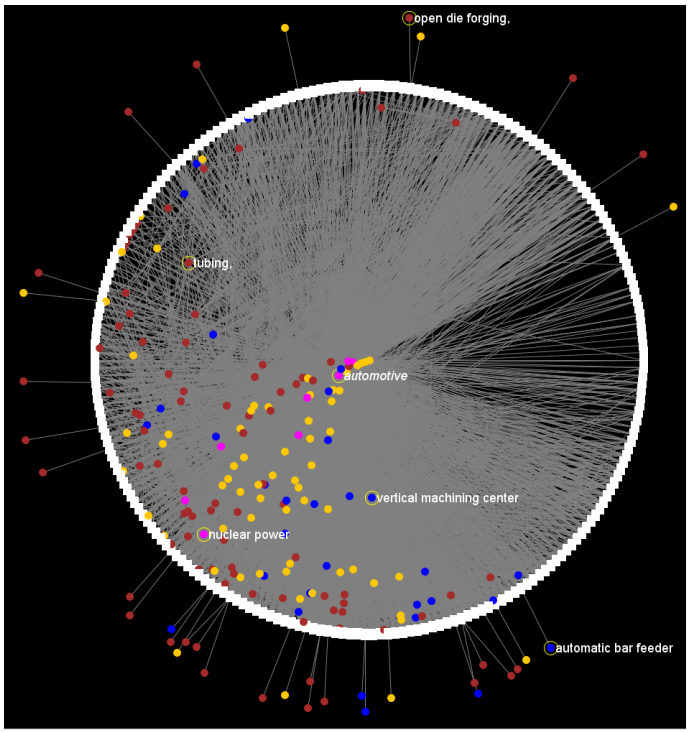


**Fig. 8:** A different arrangement of nodes in the graph view. Documents are arranged in a circular layout with all other entities located around the circle. Entities with more connections are located closer to the center.

## 5 Conclusions

This paper presented the preliminary results of using visual analytics tools for summarizing and analyzing manufacturing capability text. The Entity Extractor Tool was developed to automate the text extraction and document vectorization process. The MC Thesaurus was used to provide the Jigsaw model with the required dictionaries of entities. Only four entities were included in the experimental evaluation but in the future experiments, more entities will be included. Based on this evaluation, Jigsaw can be effectively used for capability visualization. It is easy to get started with Jigsaw but there is a learning curve involved when attempting to effectively use its multiple views in coordination. More in-depth experiments will be conducted in the future based on a larger data set to extract useful capability information from the text. Also, the EET will be integrated with Jigsaw to provide a more reusable pipeline for various text analytics and visualization purposes.

**References:**


1. Teece, D.J., G. Pisano, and A. Shuen, *Dynamic capabilities and strategic management.* Strategic Management Journal, 1997. **18**(7): p. 509-533.
2. Yazdizadeh, P. and F. Ameri. *A text mining technique for manufacturing supplier classification*. in *ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, IDETC/CIE 2015, August 2, 2015 - August 5, 2015*. 2015. Boston, MA, United states: American Society of Mechanical Engineers (ASME).
3. Keim, D., et al., *Visual analytics: Definition, process, and challenges.* Information Visualization: Human-Centered Issues and Perspectives, 2008. **4950**: p. 154-+.
4. Feldkamp, N., S. Bergmann, and S. Strassburger. *Visual analytics of manufacturing simulation data*. in *2015 Winter Simulation Conference (WSC), 6-9 Dec. 2015*. 2015. Piscataway, NJ, USA: IEEE.
5. Klemencic, M. and K. Skala. *3D visual analytics for quality control in engineering*. in *2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2013, May 20, 2013 - May 24, 2013*. 2013. Opatija, Croatia: IEEE Computer Society.
6. Parisot, O., et al. *Visual analytics for supporting manufacturers and distributors in online sales*. in *6th International Workshop on Enterprise Modelling and Information Systems Architectures, EMISA 2014, September 25, 2014 - September 26, 2014*. 2014. Luxembourg, Luxembourg: Gesellschaft fur Informatik (GI).
7. Zhang, L., et al. *Visual analytics for the big data era-A comparative review of state-of-the-art commercial systems.* in *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 14-19 Oct. 2012. 2012. Seattle WA, USA: IEEE.
8. Stasko, J., et al. *Jigsaw: supporting investigative analysis through interactive visualization*. in *2007 IEEE Symposium on Visual Analytics Science and Technology, 30 Oct.-1 Nov. 2007*. 2007. Piscataway, NJ, USA: IEEE.
9. *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*. 2005, National Information Standards Organization.
10. Miles, A. and S. Bechhofer, *SKOS simple knowledge organization system reference*. 2009, W3C.