

# Computer-Aided TRIZ Ideality and Level of Invention Estimation Using Natural Language Processing and Machine Learning

Christopher Adams<sup>1</sup> and Derrick Tate<sup>2</sup>

<sup>1</sup>Raytheon Company, Network Centric Systems, 6620 Chase Oaks Blvd. M/S 8519 Plano, Texas 75023 USA

<sup>2</sup>Texas Tech University, Mechanical Engineering Department, Box 41021, Lubbock, TX 79409-1021 USA

{Christopher.Adams, Derrick.Tate, [c-adams5@raytheon.com](mailto:c-adams5@raytheon.com)}

**Abstract.** Patent textual descriptions provide a wealth of information that can be used to understand the underlying design approaches that result in the generation of novel and innovative technology. This article will discuss a new approach for estimating Degree of Ideality and Level of Invention metrics from the theory of inventive problem solving (TRIZ) using patent textual information. Patent text includes information that can be used to model both the functions performed by a design and the associated costs and problems that affect a design's value. The motivation of this research is to use patent data with calculation of TRIZ metrics to help designers understand which combinations of system components and functions result in creative and innovative design solutions. This article will discuss in detail methods to estimate these TRIZ metrics using natural language processing and machine learning with the use of neural networks.

**Keywords:** TRIZ Natural Language Processing, Machine Learning, Neural Networks, Patent Analysis, Innovation and Creativity.

## 1 Introduction

Degree of Ideality is defined in TRIZ as “The benefit to cost ratio of the system or the ratio of its functionality to the sum of various costs associated with the building and functioning of the system” [1]. In addition, a design's level of invention is defined based on the type of design conflict resolved for a new invention and the number of disciplines used in resolving the conflict [2]. An approach for calculating a patent's degree of ideality and level of invention from patent data can be used to identify example designs that can be used as reference points during early phases of the design process to support design functional modeling and concept generation [3, 4]. This paper will discuss a computer-aided approach for extracting design functional and physical information from patent data. This approach will be used to generate hierarchical functional and physical models that are utilized to estimate TRIZ metrics. The first section includes an overview of the use of natural language processing of

patent data to extract design information from patents. The second section of the paper describes the use of patent design information to estimate the degree of ideality for each patent. The third section of the paper provides a discussion of how patent originality measures [5], number of backward patent citations made, number of forward patent citations received and the mean forward and backward citation lag can be used as training data to classify patents into the five levels of invention using machine learning techniques. Finally, the last section of the paper concludes by providing a discussion of how TRIZ metrics such as degree of ideality and level of invention can be used to support design concept generation and functional modeling during early phases of the design process.

## 2 Natural Language Processing of Patent Data

USPTO patent documents provide a good representation of a design that includes the functions performed by the design as well as the patented design's components. USPTO patent documents must follow a strict set of rules that define how a patent document is constructed [6]. Patent document rules describe different sections that must be included in the text of a patent including the patent title, abstract, claims, and patent description. Each of the patent document sections includes useful information that can be used to build functional hierarchical models for a patented invention. To extract patent functional and physical component information from patent textual descriptions, it is necessary to employ Natural Language Processing (NLP) techniques [7]. Each patent in the USPTO patent database includes drawings that depict a numbered list of all of the components of the patented design. A list of patent physical components can be used to create hierarchical functional and physical models of a patented design. This is accomplished by first extracting Subject Action Object (SAO) phrases from patent text including the title, abstract, claims, and description to retrieve each component name and number residing in the patent as well as the action verb and object that reflects the function performed by the physical components.

A number of open source Natural Language Processors were reviewed to perform the extraction of Subject Action and Objects (SAO) from the sentences in patent textual descriptions [8-11]. As part of this research the NLP software MontyLingua [11] implemented in Python was selected to perform the SAO extraction. Specialized software, the Patent SW Toolset, was generated in the Visual Basic programming environment to extract patent information from the USPTO.gov website. This software is used in the data processing phase to extract patent textual descriptions in html format. In addition, the Patent SW Toolset is used to convert patent textual descriptions previously extracted in html format into tagged XML. Tagged XML is created by the Patent SW Toolset to label and segregate the patent text so that the MontyLingua NLP software [11] can be implemented on different sections. An example of the XML report is provided below:

```
<?xml version="1.0" ?>
- <patentdata>
```

```
<patentnumber>3858357</patentnumber>

<Patent_Title>BUMPER MOUNTING FOR FLEXIBLE TRAFFIC
DOOR</Patent_Title>

  <Abstract>A door having intermediate the ends
thereof ...

...said bumpers.</Abstract>

<Inventors>McGuire; Winston B. (Plattsburgh,
NY)</Inventors>

<Assignee>W. B. McGuire Co., Inc. (Hudson,
NY)</Assignee>

<Filed_date>June 5, 1973</Filed_date>

<Current_US_Class>49/460 ; 16/86R;
160/354</Current_US_Class>

<Field_of_Search>49/460,9,34 160/354
16/86R,86A,86B,1,111R,DIG.2 293/62 85/45 29/526
42/74</Field_of_Search>

- <Citations_made>
- <citations>

<citation>3091818 June 1963 Clark</citation>

</citations>

</Citations_made>

<Claims>I claim: 1. A door having intermediate the ends
thereof a pair ...Description</Claims>

<Description>The present invention relates to doors
particularly flexible traffic doors... </Description>
```

Once XML reports are generated, the MontyLingua NLP software is used to extract the SAO instances from the different sections of the patent text. Next, an algorithm is employed in the Patent SW Toolset to extract component names by the component numbers that appear in the SAO instances that are extracted by MontyLingua from patent textual descriptions. MontyLingua provides an SAO report that includes the verb, subject, and objects that appear in each line of the patent text. As part of the subject and object extraction, it is possible to extract each component of

the patent by looking for the component number and name in the subject and object in each SAO instance. This can be used to generate a list of components for each patent.

Once the component list is generated, it is possible to generate a list of functions performed by the components by extracting functions that appear in each SAO that includes a component name and number in the subject or object. The following provides an example of an SAO extraction from MontyLingua:

```
['comprise', 'door 1', 'clear plastics upper portion 2']  
['form', 'which portion 2 and 3 are', 'at bumper 4']  
['have', 'their close position', 'overlap 6']  
['close', '', 'by magnet assembly 7']  
['include', 'door 1', 'door frame 8']  
['comprise', 'door frame 8', 'top steel frame 9',  
'steel gusset 10']  
['bolt', 'which', 'to upper and lower portion 2 and 3',  
'by bolt']  
['5', 'bolt']  
['mount', 'door 1', 'on door jamb 12', 'by door shaft  
13', 'to which shaft 13']  
['bolt', 'frame 11', 'by bolt 14']  
['reinforce', '', 'rubber seal 15']  
['extend', 'rubber seal 15', 'between longitudinal  
edge', 'of side steel frame 11']
```

The terms shown in single quotes represent the ['verb', 'subject', 'object', 'object'] extracted from each sentence using MontyLingua. The component names and numbers can be constructed in a hierarchical list by reviewing action verbs such as “comprise, form, have, include, mount, etc” that indicate that certain components are subcomponents of other components [7]. Figure 1 provides an example hierarchical component list assembled by extracting the component names and verbs from MontyLingua.

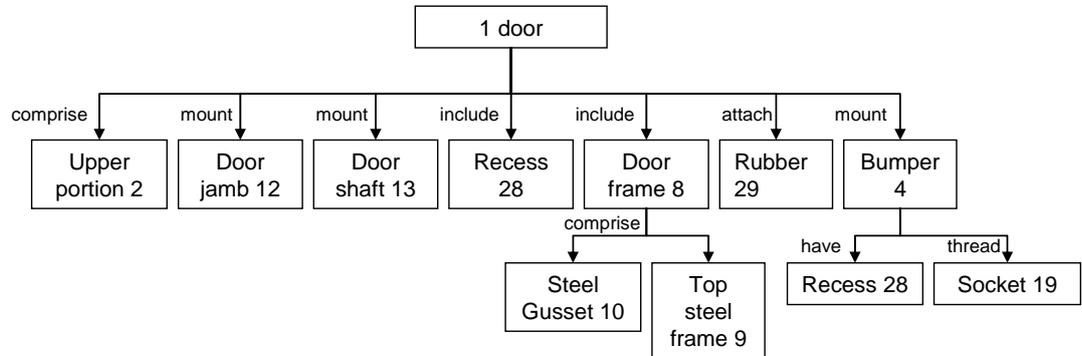


Fig. 1. Component Hierarchy for Patent #3,858,357

Once the hierarchical list of components is generated, a list of functions performed by the design can be created for each hierarchical level. This can then be used to generate patent functional models to review component and function relationships that lead to the resolution of design conflicts and increase a design's benefit to cost ratio or degree of ideality. The next section will discuss how design physical and functional models can be used to estimate a patented design's degree of ideality.

### 3 TRIZ Degree of Ideality Estimation

The process of estimating a patent's degree of ideality is conducted by first using software to automatically generate functional and physical models. Patent functional and physical models are generated with the use of patent text mining and information extraction techniques. This is accomplished by generating a list of components for each patent and modeling the functions performed by each patent component. A patent component is a physical feature, part of the system, or combination of physical parts that are used to satisfy a function fulfilled by the patented invention. Functional modeling is performed by extracting subject action object relationships for each component with the use of natural language processing techniques to perform part of speech tagging [12-14] of patent text. Patent functions and components are used to estimate patent degree of ideality by dividing the number of functions by the number of components to determine a patent's benefit to cost ratio at each functional and physical hierarchical level. For patent number 3,858,357 the degree of ideality was estimated to quantify the patented design's value. First, a list of components was generated using the method discussed in the previous section. Next, a list of functions is assembled by identifying function terms that reveal actions performed by components that represent subjects in an SAO phrase or that act on components that represent objects in the same SAO phrase. If no subject or objects exist, then the SAO phrase is not considered in the list of components. Using this approach the following list of components was extracted for patent 3,858,357:

1 Doors, 2 upper portion, 3 lower portion, 4 bumpers, 5 bolts, 6 overlap, 7 magnet assembly, 8 door frame, 9 top steel frame, 10 steel gusset, 11 side steel frame, 12 door jamb, 13 door shaft, 14 bolts, 15 rubber seal, 16 header door jamb, 17 hinge, 18 bolts, 19 threaded socket, 20 bolts, 21 threaded sockets, 22 V-cam follower, 24 recesses, 25 compression spring, 26 washer, 27 collar, 28 recess, 29 rubber

This list provides approximately 29 components for this design that represents a mount for a rubber bumper to be attached to a door. The sole function of the bumper mount is to allow the door to enable impact on the bumper. This function can be extracted from the patent text by listing all SAO phrases that include the door component name and then filtering out SAOs that do not include typical action verb terms. Once all action verbs are filtered by parsing the text and removing text other than the first term listed in single quotes it is possible to see that the only door function that remains is ['allow', 'door', 'for impact', 'on door', 'by goods'] which represents the component 'door' which allows for impact on door. (Note: Multiple objects are provided by Montylingua as part of the SAO phrase.) Other subfunctions can be found at lower subcomponent levels after SAO filtering. Functions found at subcomponent levels include thread bumper, form bumper, extend side steel frame, bolt door jamb, mount door shaft, attach V-cam follower, extend bolt, receive bolt, extend side steel frame, rise hinge, accommodate bolt, and compress spring. This represents a total of ~12 sub functions and one primary component function. By counting the subcomponents and subfunctions of the design, it is possible to estimate the degree of ideality of this system by taking the ratio of patent functions to components. This yields a ration of 13 functions to 29 components and a degree of ideality for this patented design of 45%. This metric can be used to quantify designs that provide a high level of value within a specific patent technical category. It can be used as part of the concept generation process to review designs that have a high degree of ideality and use these designs as a benchmark early in the design process. The next section will discuss using computers to estimate the level of invention for a patented design.

#### **4 TRIZ Level of Invention Estimation**

A patent's level of invention can be estimated by using patent citation analysis, patent generality and originality measures [5] as independent variables in a supervised learning model to classify patents into the five TRIZ levels of invention. The steps involved in using software to aid in the estimation of patent level of invention consist of first estimating the level of invention manually for a set of patents to use as training data for a supervised learning model. Next, data from the National Bureau of Economics Research (NBER) patent database is used as a training data set consisting of data fields such as number of citations made and number of citations received. In addition, the measures of patent generality and originality from the NBER database is used in the training set. The originality measure is calculated based on the number of patents cited by the patent under analysis that are from different patent classes. The

measure of originality is calculated using the following equation [5]:

$$o_i = 1 - \sum_{k=1}^{n_i} \left( \frac{b_{ik}}{b_i} \right)^2 \quad (1)$$

Where  $i$  is the patent under consideration,  $b$  is the number of patents cited and  $k$  indicates the subclass of the CITED patent as indicated in the NBER database. For example if one patent cites 3 patents and 2 of the patents are from subclass X and 1 patent is from subclass Y, then the originality measure is  $1 - ((2/3)^2 + (1/3)^2) = 0.44$ . A patent's generality is measured in a similar way, but considers forward patent citations to different patents from different subclasses. The measure of generality is calculated using the following equation [5]:

$$g_i = 1 - \sum_{k=1}^{n_i} \left( \frac{f_{ik}}{f_i} \right)^2 \quad (2)$$

Where  $f$  is the number of patents citing patent  $i$  and  $k$  indicates the subclass of the cited patent as indicated in the NBER database. Finally, citation information such as the mean forward citation lag and mean backward citation lag is also used as part of the network training data. This data is used to determine the breadth of influence a patented design has on future inventions.

Table 1. includes an example of the training data set used to classify patents by level of invention. It includes the independent variables as well as dependent variable, level of invention. This training data can be used with a number of different machine learning techniques to perform data classification. The machine learning technique used to perform the classification in this example is the Neural Network back propagation algorithm supplied in Matlab Neural Network Toolkit. (Other machine learning that can be used include support vector machines and Naïve Bayes Networks.) The neural network is used to train a classification model using an expanded set of training data, similar to the example training data shown in Table 1, to estimate the level of invention for a large number of patents. The patents were initially selected using the number of citations received as an indicator of patent level of invention. Intuitively, it is expected that patents that receive a large number of citations will be inventive and likely to spawn many other inventions. In addition, patents with a high forward to backward citation ratio were selected for review to identify whether these patents have a high level of invention. In addition, patents were only selected from the patents that fall in the category of mechanical designs. Mechanical designs were selected to keep the neural network from artificially selecting patents from fields such as biotechnology that draw from many diverse disciplines. It was found that using patents from many diverse disciplines skews the number of level five patent estimates.

**Table 1.** Example Training Data

patent	cmade	creceive	general	original	fwdaplag	bckgtlag	LOI
4387297	12	233	0.34	0.72	12.52	6.67	2
4575330	18	216	0.80	0.69	8.05	9.72	4
4251798	14	181	0.24	0.64	14.09	5.71	2
4409470	30	178	0.30	0.52	11.02	8.50	2
5040715	18	160	0.49	0.54	5.34	11.28	2
4835834	19	151	0.66	0.77	6.68	21.21	2
4277837	12	144	0.77	0.71	9.42	5.58	2
4361060	31	142	0.71	0.26	9.77	10.03	3
4506387	9	128	0.79	0.81	10.45	8.11	4
4834306	10	127	0.64	0.66	5.32	21.40	2
4369361	13	126	0.25	0.26	12.53	7.38	2
4520817	16	121	0.28	0.34	11.05	17.13	2
4130095	3	121	0.69	0.44	8.64	1.67	3
4944443	16	118	0.66	0.68	6.02	22.63	2
4807222	6	112	0.65	0.67	7.45	3.83	4
4728020	13	109	0.35	0.26	9.08	4.92	2
4127322	3	109	0.79	0.50	15.47	9.67	3
4236880	13	108	0.62	0.66	11.87	13.15	3
4636346	24	104	0.64	0.74	7.77	8.50	3
4303904	5	87	0.80	0.80	13.40	10.80	5
4162397	4	53	0.79	0.44	6.68	7.75	3
4714144	17	16	0.58	0.44	5.19	5.53	1
5265694	3	16	0.12	0.44	1.94	0.67	2
4074996	3	16	0.12	0.44	11.88	11.33	2
4656994	6	16	0.12	0.44	11.88	33.83	2
4717094	15	16	0.23	0.44	7.19	31.73	2
4646904	3	16	0.23	0.44	7.94	2.00	1
4051924	3	16	0.23	0.44	12.88	7.33	1
4385609	7	16	0.30	0.44	3.88	6.29	2

The patent dataset represents a larger number of mechanical and electrical patents than biotechnology and chemical patents. From reviewing the neural network results it was found that biotechnology and chemical patents typically have higher originality and generality scores than other patents. Based on this result it was determined to use only mechanical designs to train the neural network. Future research will investigate an assumption that patents that rely on knowledge across many disciplines will result in designs with higher levels of invention.

Patents with a low number of citations made and high number of citations received may be based on a new technological discovery that initiates the evolution of novel technological trends. Patents that receive a large number of citations support the evolutionary development of new technologies since it is evident that a large number of inventions result from these novel concepts. Using this training data to estimate level of invention information can also be used to understand trends in design evolution and innovation. This will aid in understanding whether TRIZ level of invention is correlated with other innovation metrics such as the emergence of a dominant design [15] and the evolution of technological discontinuities [16]. Other

potential innovation metrics are discussed as part of ongoing doctoral thesis research in [17].

A matrix data set of 43 patents was used to train the neural network back propagation model using Matlab to estimate the level of invention for a set of 48,986 patents. 23% of the patents in the data represent level one inventions, 51% of the data represents level two, 14% level three, 7% level four and 5% for level five. To train the neural network the data set was broken up into a training set that consists of 80% of the data, a validation set representing 10% of the data and another 10% of the data to test the network performance. The network was built using 100 hidden neurons to train the inputs to the targets. The inputs to the network represent the number of citations made, citations received, patent generality, patent originality, mean forward citation time lag and mean backward citation time lag for each patent. The target data represents the estimated level of invention for each patent. The level of invention data was prepared by reviewing the data set manually to make an initial estimate of level of invention for each patent. Figure 2 provides the regression performance of the training data to the neural network.

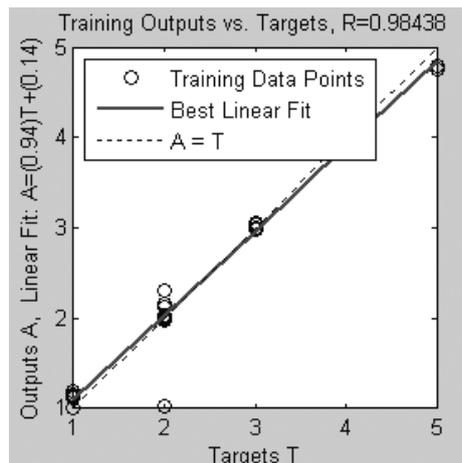
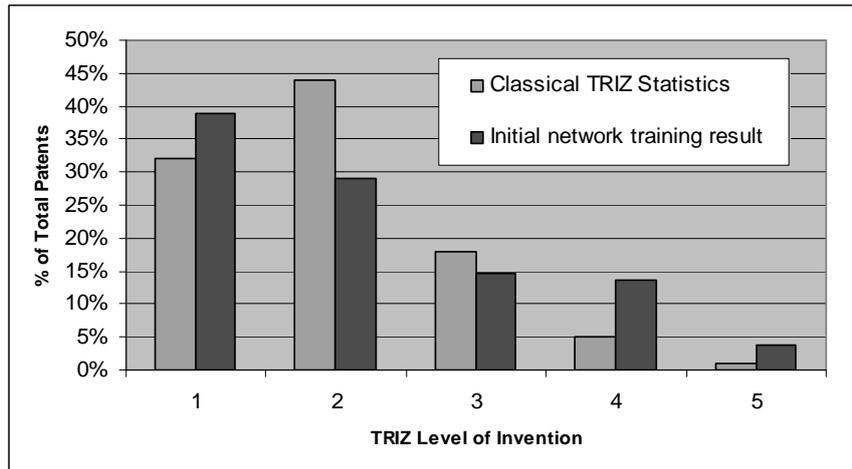


Fig. 2. Neural Network Performance Data

The figure portrays how the outputs of the neural network model, shown on the Y axis in the Figure 2 fit the target level of invention data shown on the X axis. The network fits the data well with a regression coefficient of 98.4%. This shows the network performs well based on the training data input against the level of invention target data. The neural network is then saved to the workspace to use to estimate the level of invention for a new set of patents. This new set of patents is selected using the NBER patent database by first gathering patents from the category of mechanical designs. These patents were then narrowed to patents that have a minimum of one citation made, one citation received, a measure of patent generality and patent originality in the NBER database. The input data was then run through the neural

network to classify 48,986 patents into the five levels of invention. Figure 3 provides a summary of the initial network training results from applying the network to 48,986 patents. In addition, statistics from classical TRIZ research are included for comparison to the level of invention estimates.



**Fig. 3.** TRIZ Level of Invention Statistics

The initial network results show far more patents classified as level one than the classical TRIZ research. In addition, the estimate shows far fewer patents classified as level two, close to the same number of patents for level three and far more patents estimated as level four and five. The neural network classification results fit more closely with the distribution of levels of invention described in [18]. However, the percentage of level one inventions is much higher and the percentage of level three through five inventions is much lower. The number of level three through five inventions in the distribution of level of invention data is <4% for the top three level of invention classifications. The level five patents predicted by the network were reviewed to determine estimate accuracy. One of the level five patents predicted by the network includes patent number 4,863,655 and is titled “Biodegradable packaging material and the method of preparation thereof”. This patent fits the criteria of a level four to five invention given that it represents the use of a new scientific discovery. The discovery in this patent is the use of materials for packaging materials that will degrade in the environment to prevent future environmental pollution. The neural network predicted this patent as a level five invention since it has a high citation received to citation made ratio and has a large mean forward citation time lag. This reflects that the patent has an affect over a wide range of future inventions.

Another patent identified by the neural network as a level five invention is patent number 5,232,243 titled “Occupant sensing apparatus.” This patent describes a novel method for determining when an occupant has entered a vehicle and then adjusting the seat restraint based on the size of the occupant. This involves using a material

containing an electrical characteristic that adjusts based on the size of the individual. This patent is closer to a level four invention than a level five invention. This invention is considered a level four instead of a level five invention because it lies outside of the existing paradigm of mechanical design by using a material to sense an occupant. The use of the material that contains an electrical characteristic identifies that a tool from science is used for the new invention which meets the definition of a level four invention. This patent is not a level five invention since it does not result in a new scientific discovery.

Further research will investigate methods for improving the performance of the machine learning model. One method to improve the neural network performance includes increasing the training data set by manually estimating the level of invention for more patents and increasing the number of independent variables used in the training data set. Another approach for increasing neural network performance is to base the neural network on functional and physical terms using latent semantic analysis to identify key design terms that represent each of the patents within the field of mechanical design. This method could provide a predictive measure that indicates the level of invention for more recent patents that have yet received many citations. This data will be included as part of future research to determine how neural network performance can be improved. Other areas to increase network performance may include using term frequency metrics based on the use of natural language processing and latent semantic analysis. Latent semantic analysis could be used to identify key terms that when used result in a higher level of invention.

## 5 Use of TRIZ Metrics to Support Design Innovation

TRIZ metrics such as degree of ideality and level of invention can be used early in the design process to support concept generation, functional modeling, and functional synthesis. A wide range of patents that span across multiple disciplines can be analyzed to see how new design components evolve over time to perform existing system functions or create new functions by integrating cross domain knowledge. Patent citation networks are used to identify patents that initiate technological discontinuities by using a new set of components to perform common functions. These trends of evolution can be examined using metrics such as degree of ideality and level of invention. In addition, new metrics can be created by employing text mining techniques to extract a set of key words that represent a given discipline. These key words identify how components from multiple disciplinary fields can be integrated to develop new technologies to perform existing system functions. [17] Data mining techniques can be used to generate TRIZ metrics for a large number of patents. Machine learning techniques can then be used to train neural network, or other machine learning, models to uncover evolutionary trends that reside in patent data. These trends can be used in the concept generation process to review design trends that lead to innovative design concepts. As part of future research, the use of machine learning techniques and approaches will be used to train prediction models that can help in the innovation process by predicting future success of new designs.

## References

1. Fey, V., Rivin, E.: *Innovation on Demand: New Product Development Using TRIZ*: Cambridge University Press, (2005)
2. Altshuller, G.: *Creativity as an Exact Science*: CRC Press, (1994)
3. Regazzoni, D., Nani, R.: "TRIZ-Based Patent Investigation by Evaluating Inventiveness," in *IFIP International Federation for Information Processing, Volume 277; Computer-Aided Innovation (CAI)*; Gaetano Cascini; (Boston: Springer), pp. 247-258 (2008)
4. Verbitsky, M.: "Semantic TRIZ," Invention Machine Corporation, Boston (2004)
5. Jaffe, A.B., Trajtenberg, M. *Patents: Citations & Innovations A Window on the Knowledge Economy*: MIT Press, (2002)
6. USPTO, "Patent Rules: Title 37 - Code of Federal Regulations - Patents, Trademarks, and Copyrights." vol. Title 37, (2000)
7. Fantechi, A., Cascini, G., Spinicci, E.: *Natural Language Processing of Patents and Technical Documentation*. Berlin Heidelberg, (2004)
8. Hendrickson, C., Charniak, E., Jacobson, N., Perkowitz, M.: "Equations for Part of Speech Tagging," in *In Proceedings of the Eleventh National Conference on Artificial Intelligence*, Menlo Park, pp. 784-789. (1993)
9. Caroll, G., Charniak, E. Adcock, et al.: "Taggers for Parsers," *Artificial Intelligence*, vol. 85, (1995)
10. Liu, H.: "ConceptNet: A Practical Commonsense Reasoning Toolkit," *BT Technology Journal*, vol. 22, pp. 211-226, (2004)
11. Liu, H.: "MontyLIngua website, ([web.media.mit.edu/~hugo/montylingua/](http://web.media.mit.edu/~hugo/montylingua/))," (2004)
12. Brill, E.: "A simple rule-based part of speech tagger," in *In proceedings of the Third Conference on Applied Natural Language Processing*, (1992)
13. Brill, E.: "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging," *Computational Linguistics*, vol. 21, pp. 543-565, (1995)
14. Charniak, E.: "Statistical techniques for natural language parsing," in *AI magazine*. vol. 18, pp. 33-43. (1997)
15. Suarez, F.F., Utterback, J.M.: "Dominant Designs and the Survival of Firms," *Strategic Management Journal*, vol. 16, pp. 425-430, (1995)
16. Christensen, C.M.: *The Innovator's Solution: Creating and Sustaining Successful Growth*, (2003)
17. Adams, C.: "Developing Transdisciplinary Metrics Using Data Mining Techniques," in *Mechanical Engineering Department Lubbock, TX: Texas Tech University*, Expected Spring (2009)
18. Zlotin, B. Zusman, A.: "Levels of Invention and Intellectual Property Strategies," *Ideation International*, Southfield, Michigan (2003)