Chapter 11
Methodological Demonstration of a Text Analytics Approach to Country Logistics System Assessments

Aseem Kinra, Raghava Rao Mukkamala and Ravi Vatrapu

Abstract The purpose of this study is to develop and demonstrate a semi-automated text analytics approach for the identification and categorization of information that can be used for country logistics assessments. In this paper, we develop the methodology on a set of documents for 21 countries using machine learning techniques while controlling both for 4 different time periods in the world FDI trends, and the different geographic and economic country affiliations. We report illustrative findings followed by a presentation of the separation of concerns/division of labor between the domain expert and the text analyst. Implications are discussed and future work is outlined.

Keywords Logistics · Transport system evaluation · Big data analytics · Data mining · Text mining · Machine learning

A. Kinra (✉)
Department of Operations Management, Copenhagen Business School, Solbjerg Plads 3, 2000 Frederiksberg, Denmark
e-mail: aki.om@cbs.dk

R.R. Mukkamala · R. Vatrapu
Computational Social Science Laboratory, Department of IT Management, Copenhagen Business School, Howitzvej 60, 2000 Frederiksberg, Denmark
e-mail: rrm.itm@cbs.dk

R. Vatrapu
e-mail: rv.itm@cbs.dk

R. Vatrapu
Mobile Technology Laboratory, Faculty of Technology, Westerdals Oslo ACT, Oslo, Norway

M. Freitag et al. (eds.), *Dynamics in Logistics*, Lecture Notes in Logistics,
DOI 10.1007/978-3-319-45117-6_11

11.1 Introduction

The assessment and appraisal of transportation systems is a central activity related to both national and regional policy making for passenger and freight movement. In this regard, there are a fixed set of widely accepted core methods for economic evaluation and appraisal for different project investments, e.g., related to transport infrastructure for the movement of people and goods. Most of these methods that are currently used for this purpose involve the calculation of benefit-cost ratios.

There is an inherent assumption in that such an evaluation and assessment process will aid economic goals, and even foster development in the case of developing countries by lowering the overall transportation and mobility costs of businesses and private users. However, the cross-border movement of goods involves international transportation, global supply chains, multinational businesses and users, where the costs and benefits of transportation systems are not only highly variable, but also tied to the broader global logistics and supply chain systems that embed these transportation systems. The Multi National Corporation (MNE), and its global supply chain then becomes a primary user and motor of economic growth in an era of globalization and global integration of businesses. Tracking the business considerations of MNEs, and how their assessments of the national and regional transportation, logistics and supply chain systems play a role in these considerations, becomes important for policy makers in order to increase international trade and to attract foreign direct investments in almost all sectors of the economy.

Toward this end, this paper explores and exploits text analytics methods and techniques from big data analytics to develop a methodology for information scanning, extraction, and retrieval of the logistics-related measures preferred by decision-makers of MNCs, and other firms involved in global supply chains. This information is relevant for the benchmarking and evaluation of transportation and logistics systems. We use text mining and text analytic approaches, and more specifically machine learning techniques in order to develop this methodology. We then test run this methodology on a global supply chain text corpus in order to illustrate and provide some feedback on the methodology. Our initial results illustrate that the methodology is rather useful, and it can be successfully applied to other logistics and transportation applications. However, we can also relate to the general observation that model specification is the Achilles' Heel of big data analytics as it is easy to under- or over-specify the model without deep involvement of domain experts and deep knowledge of domain-specific problems.

11.2 The Assessment of Country Logistics Systems

The assessment of country logistics systems is an emerging area and literature is sparse. Though most of the existing work on the assessment of country logistics systems can be related to the useful distinctions in existing location research (see

Beugelsdjik et al. 2010; Beugelsdjik and Mudambi 2013). Kinra (2015) traces the main literature streams in this area to the international economics, economic geography, and international business distinctions in existing location research, and these literature streams are now summarized.

The economics-oriented literature generally adopts the stance that spatial variation is generated by the structure of national resource endowments, and trading patterns. The main contributions in this stream come from Memedovic et al. (2008) and Bookbinder and Tan (2003), among others. For this group of literature spatial discontinuity generally occurs at national borders and is defined by national logistics systems. The geography-oriented literature tends to adopt the stance that connectivity and colocation effects at major trade interface points generate spatial variation. The contributions of Rodrigue and Notteboom (2010) and Rodrigue (2012) are prominent in this stream. For these contributors spatial discontinuity occurs at major nodes: transit points, hubs, and gateways, and is defined by regional logistics systems, and hence the focus on the ranking of regional logistics systems. Finally, the international business-oriented literature tends to adopt the stance that the managerial and organizational utility from locations generates spatial variation. Min (1994) is one of the earliest to demonstrate this stance. Similarly Carter et al. (1997), Kinra and Kotzab (2008) make important advances in this regard. For this group of contributors, spatial discontinuity generally occurs by country, and is defined by the country investment climate.

All these contributions seek to provide an evaluation of the national logistics systems. However some issues are common in the literature. First, the literature is not unanimous about the logistics-oriented determinants, or factors, of location. Second, it does not clarify the role and type of information, or measures, that are adopted in the assessments. The literature is therefore mute about the specific spatial impedance information and information categories that are relevant for decision-makers in country assessment, and it is indefinite about the relevance and the importance of the considered location determinants and the missing information. This is a problem because information about the foreign country is widely recognized as a major contributor in the complexity under conditions of globalization (see McCann and Mudambi 2004), and global supply chains.

We make a first attempt to the identification and categorization of the relevant, available information on country logistics systems in Kinra (2015). Apart from the identification of 187 different types of relevant and essential information measures that fall under 17 decision factors, the study establishes the main spatial transaction cost categories under which the information is made available, but also concludes on conditions of information-driven complexity for the ranking of country logistics systems.

However these findings are only based only complex manual procedures for discerning this information, and include observations from a limited dataset of countries for a limited time period (2006–2008). Nor any implications are presented for the ranking (assessment) of the different countries based on the spatial impedance measures that have been identified. This study will go beyond these

limitations by developing more automated techniques that are capable of handling bigger information sets. This will take better into account factors such as longitudinality of data for determining the general conditions of information-driven complexity hypothesis, but also from the point of view of comparing different time periods. The next section now provides a detailed description of this methodology development, including the text mining and text analytic approaches that are employed to develop the methodology for automating the information scanning, extraction, and retrieval process.

11.3 Methodology

In this section, we describe the methodology that was developed to conduct big data analysis of global supply chain document corpus. The corpus primarily contains 21 text documents describing transportation and logistics systems pertaining to 20 countries spanning over period from 2006–2014 years. In order to analyze the text corpus, we have used both text mining and machine learning techniques as shown in Fig. 11.1.

As a first step (Text Extraction) toward applying text mining/machine learning algorithms on the text corpus, we have extracted texts from the global perspective documents which are in Portable Document Format (PDF) using an open source software component using Java Programming language. The size of a global

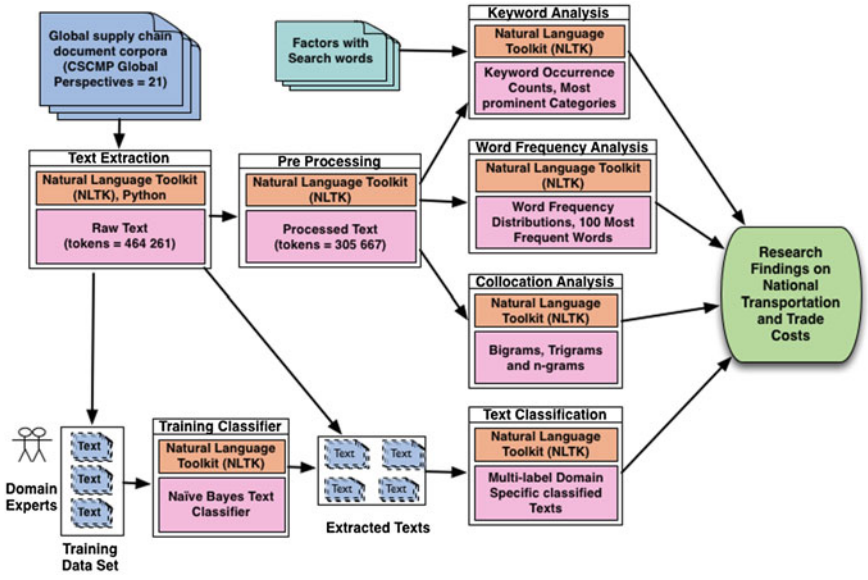

Fig. 11.1 Text analytics methodology of global supply chain document corpus

perspective document varies from 20 to 60 pages and text from each page has been extracted separately for further processing. Unfortunately, it was not possible for us to extract the PDF metadata such as links and bookmarks from the documents. The extracted texts are tokenized by breaking them sentences and words by using space as delimiting character between the words using Natural Language Toolkit (NLTK) (Bird 2006) and Python programming language. The final output of this step ended up in raw text containing approximately 464,000 tokens, which is used as an input for further processing using text mining and machine learning algorithms.

11.3.1 Text Mining

As part of text mining we have applied three different techniques: *Keyword Analysis*, *Word Frequency Analysis*, and *Collocation Analysis* as further described below. In order to prepare the raw text for applying these techniques, we had to clean and preprocess the raw text using Natural Language toolkit. As part of preprocessing step, we further processed the raw text tokens by removing nonalphabetical characters, numerical and stop words that are high-frequency words like *the*, *to, also, and* so on, which do not contribute much to the semantic meaning of the documents. Even though the raw text contains 464 000 tokens, after the preprocessing step, we ended up with approximately 300,000 tokens.

11.4 Keyword Analysis

As part of keyword analysis, we investigated which decision factors are more predominant in the country descriptions. In order to investigate that, each factor (e.g., Waterways) is supplemented with suitable search words (port, water, sea, shipping, etc.) that are representative of that factor. Even though 17 decision factors with suitable search words are adopted from Kinra (2015) initially, soon the list of factors got expanded to 21 factors to include more interesting factors. One of the challenges with the decision factors from Kinra (2015) is that the number of search terms varied from factor to factor which made the cross comparisons of decision factors among the countries difficult. Therefore, we have readjusted search terms by dropping few search terms for some factors and by adding new search terms for some other factors, thereby making the number of search terms equal to four for every decision factor. Soon after finalizing the search terms for decision factors, we used NLTK with Python programming language to conduct analysis to find out the number of occurrence of each search term (or keyword) in each perspective document for all the decision factors. The occurrence counts of search terms are summed up to find out the keyword occurrence count for a decision factor. Based on the keyword occurrence counts of decision factors, we were able to compare which decision factors are predominant in the entire corpus and as well as in each

country descriptions. In addition to that, we could also be able to compare decision factors across country dimension such as which are the most prominent countries for any given decision factor.

11.5 Word Frequency Analysis

Word frequency analysis is a method of automatically identifying the frequent occurring words from a given text corpus, by using the term document matrix. In order to compute the word frequencies, raw tokens from the preprocessing step are further analyzed using NLTK to prepare term document matrix containing the frequency of words across perspective document collection. We have computed the term document matrix for the whole corpus as a single document and for each individual perspective document as well. The construction of term document matrices enables us to find most frequent words (e.g., 100 most frequent words) with the word frequencies, which is used to generate word cloud for a given document. Word clouds or tag clouds is a simple way of visualizing or highlighting the most frequently used words from a given text document. As part of the analysis, we have generated word clouds for each perspective document and also a word cloud for whole text corpus by combining all the perspective documents into one. Word frequency analysis and word clouds enabled us to get an overview of major concepts/topics discussed in the documents.

11.5.1 Collocation Analysis

Collocations are expressions of multiple words, which commonly co-occur in the documents and therefore a collocation is a sequence of words that occur together unusually often in the text documents. Finding collocation expressions involves standard statistics-based and linguistically rooted association measures against mere frequency of word occurrence counts. The collocation analysis provides insights about the documents by providing bigrams, trigram, and n-grams that contain words, which co-occur in the documents. The collocation analysis for each perspective document as well as for whole data corpus was conducted using Natural Language toolkit (Bird 2006) to find out bigrams and trigrams. It provided us with intuition about the topics and concepts that are discussed in the country descriptions of transportation and logistics. It also helped us notice the emerging trends (other than decision factors) in the perspective documents, which further helped us process the documents for emerging trends or factors.

11.5.2 Machine Learning and Text Classification

As part of the methodology, our plan is to apply machine learning algorithms on the text corpus to perform text classification tasks. Text classification approach comes under supervised machine learning and it can be defined as a process where assigning a predefined category of labels to new documents based on probabilistic measure of likely hood using a training set of labeled documents (Yang and Liu 1999). Out of several approaches available in text classification domain, we have chosen a simple text classification method (Zhang and Li 2007) based on Bayes rule that relies on a simple representation of documents using bag of words approach. In this project, we have used Naïve Bayes classification method to classify the extracted text pieces from the perspective documents.

First, we have extracted text portions from the perspective documents of country descriptions using NLTK based on occurrence of each search word that belong to a decision factor. In order to get proper understanding of the context of the occurrence of the search word, we have extracted few sentences of text on both sides of the occurrence location of search word. Altogether, we have extracted around 28,000 text portions from the total 21 perspectives documents. The accuracy of text classifier purely depends on the quality of manually encoded training set data. Therefore, in order to get a proper training set that is representative of the transportation and logistics domain, we are planning to get 10 % of the extracted text portions manually coded by the domain experts belong to Transportation and Logistics domain. As shown in Fig. 11.1, major part of the manually coded texts will be used for training the Naïve Bayes text classifier. After the classifier gets sufficiently trained we will use the rest of the training set to test the accuracy of the classifier. Finally, the classifier will be used to on the remaining 90 % of extracted texts to classify them automatically using the machine learning technique. Furthermore, for the text classification we would like to different domain-specific models such as sentiment, emotion, and any other models that are specific and suitable for the transportation and Logistics domain.

11.6 Illustrative Results

We present a selection of the results below to illustrate the methodological approach and to demonstrate the domain-specific utility of applying text analytics methods and techniques (Bird 2006; Yang and Liu 1999; Zhang and Li 2007). Table 11.1 presents the category analysis of the text corpus to extract the basic indicators of global supply chain perspectives.

Figure 11.2 shows the five most frequent categories and their respective cumulative keyword occurrences across the different time periods.

Table 11.1 Category analysis of global supply chain perspectives

Categories	2004–2007	2008–2009	2010–2013	2014
Waterways	1361	2232	2965	1117
Intermodal	668	701	1119	424
Economic structure	396	464	917	225
Railways	363	486	677	337
Economic policy	321	239	581	124
Airways	185	674	849	299
Geographical location	108	548	649	120

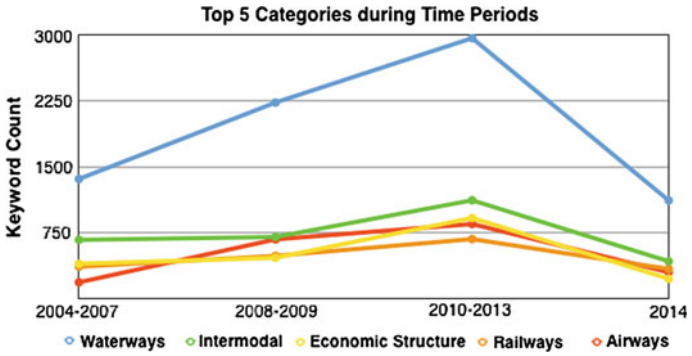

Fig. 11.2 Top 5 categories during time periods

Fig. 11.3 Top categories and measures for whole dataset

Figure 11.3 shows the most frequent categories and measures for the whole data corpus. Finally, Fig. 11.4 shows the emergent categories and measures for the whole data corpus.

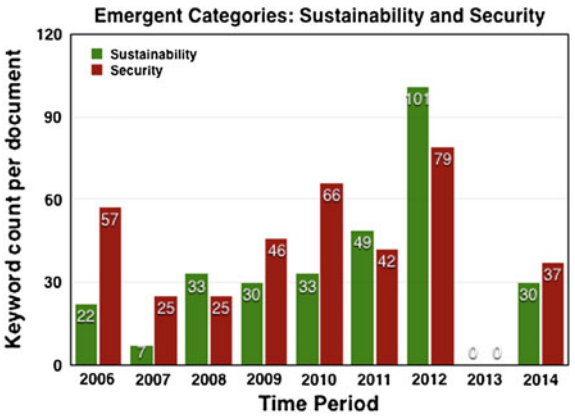

Fig. 11.4 Emergent categories of sustainability and security

11.7 Discussion and Conclusion

Figure 11.5 presents the separation of concerns/division of labor between the domain expert and the text analyst. As can be seen, the deep involvement of the domain expert during the model specification and training phases of the text analytics process is necessary but not sufficient. Subsequent involvement of the domain expert during the classification and model fine-tuning phases yields empirical results that are robust, reliable, and relevant. We will further explore the dynamics

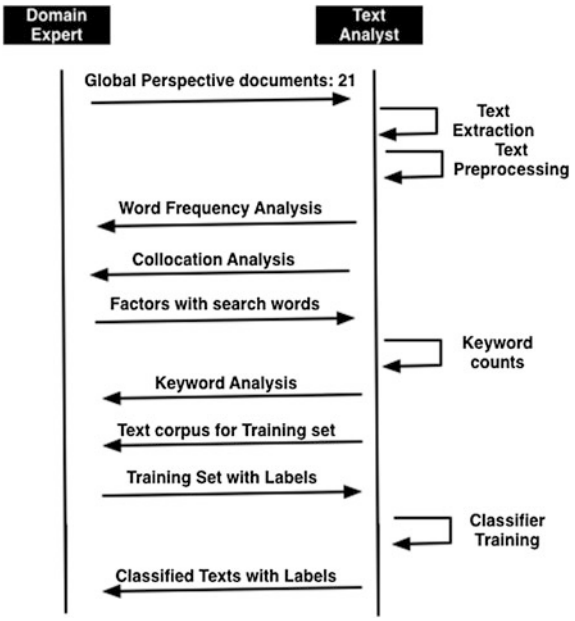

Fig. 11.5 Separation of concerns: domain expert versus big data analyst

of the relationship between specific domain expertise and generic big data analytics expertise in our future work.

This study has developed and demonstrated an approach for automating the information scanning, retrieval, and extraction process for logistics and transportation applications. The approach is especially useful because it will aid in the systematic study of big data while making sense of different types of logistics costs that are currently being incurred in a tumultuous market place. Companies are increasingly facing big data challenges in the global transportation of goods and services, and look for dedicated approaches in this regard in order to leverage their information system investments for optimizing their logistics and transportation costs. For example, Maersk Line, the world's largest container shipping company has recently implemented a "Remote Container Management" system that collects information using sensors and transceivers embedded in the containers. Such systems create a deluge of near real-time data that need to be modeled and analyzed using methods and techniques from the emerging field of data science in general and big data analytics in particular. However, we can also relate to the general observation that model specification is the Achilles' Heel of big data analytics as it is easy to under- or over-specify them without deep involvement of domain experts and deep knowledge of domain-specific problems.

In addition to traditional text mining and machine learning methods, the next steps involve employing the new method of "Social Set Analysis" to extract and model latent social information and the resulting social influence, if any, from the document corpus. Vatrapu et al. (2014a, b) have proposed a set theoretical approach to big social data analytics called Social Set Analysis (SSA). SSA is based on the sociology of associations and the mathematics of set theory and supports both interaction analytics in terms of actors involved, actions taken, artifacts engaged with as well as text analytics in terms of keywords employed, feelings expressed, pronouns used, and topics discussed (Mukkamala et al. 2014a, b; Vatrapu et al. 2014a, b). For the purposes of this paper and the domains of trade and transportation, we plan to employ SSA to uncover the temporal distribution of institutional actors' engagement as well and unique actor/keyword sets before, during, and after events of theoretical interest and the spatiotemporal dynamics of keywords and expressed feelings with regard to country assessments in the text corpus.

References

Beugelsdijk S, Mudambi R, McCann P (2010) Place, space and organization: economic geography and the multinational enterprise. J Econ Geogr 10(4):485–493

Beugelsdijk S, Mudambi R (2013) MNEs as border-crossing multi-location enterprises: The role of discontinuities in geographic space. J Int Bus Stud 44

Bird S (2006) NLTK: the natural language toolkit. Paper presented at the proceedings of the COLING/ACL on interactive presentation sessions

Bookbinder JH, Tan CS (2003) Comparison of Asian and European logistics systems. Int J Phys Distribut Logist Manag 33(1)

Carter JR, Pearson JN, Peng L (1997) Logistics barriers to international operations: the case of the people's republic of China. J Bus Logist 18(2):129–145

Kinra A, Kotzab H (2008) Understanding and measuring macro-institutional complexity of logistics systems environment. J Bus Logist 29(1):327–346

Kinra A (2015) Environmental complexity related information for the assessment of country logistics environments: implications for spatial transaction costs and foreign location attractiveness. J Transp Geogr 43:36–47

McCann P, Mudambi R (2004) The location behavior of the multinational enterprise: some analytical issues. Growth Change 35(4):491–524

Memedovic O, Ojala L, Rodrigue JP, Naula T (2008) Fuelling the global value chains: what role for logistics capabilities. Int J Technol Learn Innov Dev 1(3):353–374

Min H (1994) Location analysis of international consolidation terminals using the analytic hierarchy process. J Bus Logist 15(2):25–44

Mukkamala R, Hussain A, Vatrapu R (2014a) Fuzzy-set based sentiment analysis of big social data. In: Proceedings of IEEE 18th international enterprise distributed object computing conference (EDOC 2014), Ulm, Germany, pp 71–80. doi:1510.1109/EDOC.2014.1519. ISBN: 1541-7719/1514

Mukkamala R, Hussain A, Vatrapu R (2014b) Towards a set theoretical approach to big data analytics. In: Proceedings of the 3rd IEEE international congress on big data 2014, Anchorage, United States

Rodrigue J-P (2012) The geography of global supply chains: evidence from third-party logistics. J Supply Chain Manag 48(3):15–23

Rodrigue J-P, Notteboom T (2010) Comparative North American and European gateway logistics: the regionalism of freight distribution. J Transp Geogr 18(4):497–507

Vatrapu R, Mukkamala R, Hussain A (2014a) A set theoretical approach to big social data analytics: concepts, methods, tools, and findings. Paper presented at the computational social science workshop at the European conference on complex systems 2014

Vatrapu R, Mukkamala R, Hussain A (2014b) Towards a set theoretical approach to big social data analytics: concepts, methods, tools, and empirical findings. Paper presented at the 5th annual social media & society international conference 2014

Yang Y, Liu X (1999) A re-examination of text categorization methods. Paper presented at the proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval

Zhang H, Li D (2007) Naïve Bayes text classifier. Paper presented at the granular computing, 2007. IEEE international conference on GRC 2007