Geographic Information Systems (GIS) Based Visual Analytics Framework for Highway Project Performance Evaluation

86

Chau Le, Tuyen Le, and H. David Jeong

Abstract

Advances in data and information technologies have resulted in the availability of different types of useful data and easy accessibility. However, much of the data is not fully leveraged to gain insights for decision making due to the labor-intensive and time-consuming process of data collection and the unstructured format of data, which imposes challenges for data analytics. This paper discusses a Geographic Information Systems (GIS) based visual analytics framework for highway project performance evaluation in terms of cost and productivity. The study employs web data extraction techniques and database technologies to automatically extract data of interest from different web data sources to develop databases that contain structured data for data analytics. To merge the data from distinct sources, natural language processing techniques are also used to deal with the inconsistency of data terminology and word choices. In addition, the use of GIS technologies allows for the visualization and analysis of data collected from different locations. A case study was undertaken, which implemented part of the framework for unit price visualization, estimation, and evaluation of highway projects.

Keywords

Visual data analytics • Web data extraction • Database • Natural language processing
Geographic information systems • Unit price predictions and visualizations

86.1 Introduction

Advances in data and information technologies have resulted in the availability of different types of useful data and easy accessibility. However, much of the data is not fully leveraged to gain insights for decision making due to the labor-intensive and time-consuming process of data collection and the unstructured format of data. In the highway sector, State Departments of Transportations (DOTs) have been collecting data about highway projects and publishing the data on their websites for public view. The data includes a variety of information for each project such as project plans, location, schedule, scopes of work, and estimated costs; but most of them are in unstructured formats (e.g., PDF files). There is a need for an efficient way that can extract useful information from DOT's websites and integrate different types of information for data analytics to gain understanding of highway projects.

This paper discusses a Geographic Information Systems (GIS) based visual analytics framework for highway project performance evaluation regarding cost and productivity. The study employs web data extraction techniques and database

C. Le (✉) · T. Le · H. D. Jeong
Department of Civil, Construction and Environmental Engineering, Iowa State University, Ames, IA 50011, USA
e-mail: chle@iastate.edu

T. Le
e-mail: ttle@iastate.edu

H. D. Jeong
e-mail: djeong@iastate.edu

I. Mutis and T. Hartmann (eds.), *Advances in Informatics and Computing in Civil and Construction Engineering*,
https://doi.org/10.1007/978-3-030-00220-6_86

technologies to automatically extract data of interest from different web data sources to develop databases that contain structured data for data analytics. To merge the data from distinct sources, natural language processing techniques are used to deal with the inconsistency of data terminology and word choices. Also, the use of GIS technologies allows for the visualization and analysis of data collected from different locations. A case study was undertaken, which implemented part of the framework for unit price visualization, estimation, and evaluation of highway projects.

86.2 Background

86.2.1 Web Data Extraction Techniques

The internet contains a massive amount of data from various sources and in different formats. Extracting data from web pages and transforming it into structured forms are necessary for data analysis to explore useful information and gain understanding about phenomena of interest [4, 11]. Web data extraction techniques come from a variety of areas and disciplines, such as natural language processing, machine learning, databases, and ontologies [21]. The techniques can be divided into two main approaches: tree matching and machine learning [11, 31]. The former is based on the semi-structured format of the HTML web pages to identify the location of the required information. The latter learns from examples labeled by domain experts to extract data from similar unseen websites by using machine learning algorithms such as convolutional neural networks [12, 31].

86.2.2 Natural Language Processing

Natural Language Processing (NLP) is an area of artificial intelligence which includes a collection of techniques that can process, analyze, and manipulate human language data such as text and speech. Some of the techniques are tokenization [29, 32], Part-of-Speech tagging [7, 27], and Named Entity Recognition [1]. NLP research has been around since the 1950s and has been applied in various tasks, for example in translation, information extraction, information retrieval, and topic modeling [3]. These applications are being supported by the availability of highly accurate NLP packages and libraries such as NLTK and Gensim.

Information Extraction. Information extraction is the process of extracting desired structured information from text documents. It includes two major tasks name entity recognition (NER) and relation extraction (RE) [19]. NER aims to find names of entities such as organizations (e.g., "Iowa State University"), persons (e.g., "Joanna Shaffer"), places (e.g., "Ames, Iowa"), time (e.g., "April, 2018"), and numbers (e.g., "11"). RE refers to finding relationships between entities, for example: "Iowa State University" is located in "Ames, Iowa" [26]. Two main approaches for NER are the rule-based approach and the statistical learning approach. Rule-based methods use a set of predefined rules to find matches in text documents. Those rules can be manually developed by domain experts or automatically identified through applications of machine learning to NLP [18, 24]. The statistical learning approach treats the text as a sequence of words with labels, and each one depends on the others in the series [19].

Semantic Measures. One of the main NLP related research topics is semantic measurement, which aims to evaluate the similarity or relatedness between semantic units (words, phrases, sentences, concepts, etc.) [15]. Semantic similarity indicates the likeness in the meaning of different semantic units; for instance, the words "road" and "highway" are semantically similar even though they do not share the same string representations. The two primary approaches for semantic measures are (1) knowledge-based method and (2) corpus-based method [15]. The former builds upon ontologies or digital dictionaries, which are lexical networks of terms and their semantic relations such as synonym and hypernym. The relatedness of the two words is measured by a similarity measure such as the distance between them in the network. However, ontologies and digital dictionaries are not available to many domains. The latter method relies on corpora of text and distributional models from those corpora. A distributional model stands on the distributional hypothesis: two words that occur in the same contexts have similar meanings [16]. Therefore, distributional models represent a word through its surrounding words observed from a given corpus [10]. The outcome of this approach is a Vector Space Model (VSM), where each word is represented by a point in a high-dimensional space. Two words with similar contexts are close to each other in the vector space.

A considerable amount of literature has been published on learning vector representations of words by using neural networks [23, 25, 28]. For example [23] proposed two neural network models (also known as word2vec): continuous

bag-of-word model (CBOW) and skip-gram model. For the CBOW model, the target word is at the output layer, and context words (surrounding words) are at the input layer. Conversely, for the skip-gram model, the target word is at the input layer, and the context words are at the output layer. Also, several attempts have been made to learn distributed representations of phrases, sentences, paragraphs, and documents by extending word representations in vector space [13, 22]. Paragraph vector (also known as doc2vec) was proposed by Le and Mikolov [22] is an extension of word2vec, which is capable of constructing vector representations of word sequences of various length (sentences, paragraphs, and documents).

86.2.3 Geographic Information System and Spatial Interpolation Methods

GIS technologies have been increasingly used to handle spatial data in various application areas, such as urban planning, transportation planning, and environmental management. It allows for positioning real properties or objects on a local map based on their geographical coordinates [8]. GIS represents some aspect of the real world through spatial models using simplified spatial entities like points, lines, and areas [17]. Each entity is associated with one or more attributes that give additional information about that entity [17]. By organizing, integrating, analyzing different types of data, GIS can handle complex spatial data and create new information to support decision making [6, 17].

Interpolation techniques is an essential component of GIS [9]. Spatial interpolation, as defined by Burrough et al. [2], is the prediction of the value of a variable at unknown locations within the area surrounded by known data points. Spatial interpolation techniques can be considered to consist of two categories: deterministic ones (e.g., inverse distance weighted and global polynomial) and geo-statistical ones (e.g., ordinary kriging, simple kriging, and cokriging) [5, 20]. Deterministic interpolation depends on the values of measured points or mathematical calculations from them while geostatistical methods rely on statistics (e.g., spatial autocorrelation among the measured locations). Some popular interpolation techniques are supported by GIS platforms. One of the most popular ones is ArcGIS which support spatial data analysis to create continuous surfaces or maps from measured data, along with errors of the predicted surfaces [20].

86.3 GIS-Based Visual Analytics Framework for Highway Project Performance Evaluation

This paper proposes a GIS-based visual analytics framework for highway project performance evaluation. The idea arises from applications of GIS in other research areas such as analysis of heavy metal sources in soil [14] or analysis of groundwater level in a specific region [30]. In this proposed framework, each highway project can be modeled as a spatial entity such as a point or a line on a map. In case that the purpose of a study is to make comparisons among many projects in a large area, point entities can be used to present highway projects. Each point can be associated with different types of data: spatial data (e.g., latitudes and longitudes), temporal data (e.g., recorded date of the data), and other un-spatial data that contain specific information about the project. Each project has its characteristics, and many of them are influenced by project locations, such as weather, topographical conditions, and geotechnical conditions. By integrating different types of data, researchers not only can implement standard analyses for un-spatial data but also explore relationships among variables or influences of time and location on other characteristics of highway projects. Thanks to GIS, spatial questions relating to highway projects can be answered to support decision-making processes. For example, developing a unit price map of a work item over several states can help contractors see some overall patterns or variations of unit prices in different states, from which they can have appropriate bidding strategies for each state. By such comparisons, performances of a project can be evaluated through cost, productivity, and performance indexes when those of other similar projects are known and shown on the same map.

A next question that needs to be addressed is how to obtain those kinds of data of highway projects for data analysis. Ideally, data of interest can be provided by data creators such as DOTs. Those ideal situations are not common and the corresponding processes still heavily rely on human interactions. There is, however, another option. With the availability of data in DOTs' websites, the desired information can be automatically harvested using web data extraction and NLP information extraction techniques. For example, an algorithm can be developed to obtain bid tabulations from DOTs' websites and then extract desired information (e.g., bid item code, item description, unit, unit price, and location) through NER tasks. To evaluate the accuracy of the algorithm, a testing dataset, which is developed by domain experts, is needed to compare the output of the algorithm with the one identified by the experts. In addition, outlier analysis may be employed to exclude information that is extracted by errors. After that, data is ready for analysis of projects within the same state. For comparisons across different states, the issue of inconsistencies in data terminology and word choices also needs to be dealt

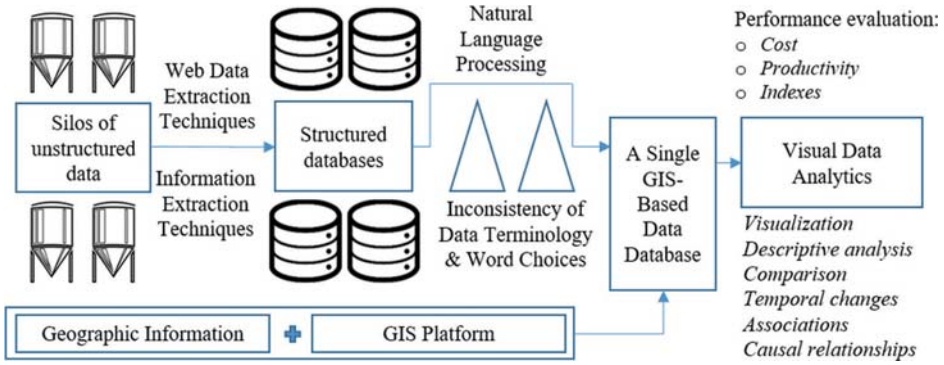

Fig. 86.1 A GIS-based visual analytics framework

with because each DOT has their own specifications, standards, and work breakdown structures. To overcome this barrier, manual matching of variable descriptions by experts is possible, but time-consuming and labor intensive. To save time and reduce human involvement in the process, NLP techniques can be utilized to automatically measure semantic similarity between word units. The results from automatic matching should then be evaluated by domain experts to ensure the reliability of the combined data.

The above discussions are organized and summarized to form a GIS-based visual analytics framework as Fig. 86.1.

The following is a case study which partially implemented the proposed framework. The authors collected bid data for the year 2013 from Iowa DOT and Montana DOT. The bid items selected for illustration were "Asphalt binder, PG 64-28" (Iowa DOT, item code "2303-0246428") and "Asphalt cement, PG 64-28" (Montana DOT, item code "402020092"). With the assumption that the two bid items represent the same work, unit prices of 17 projects in Iowa and 26 projects in Montana were used for comparisons. By using an interpolation method available in ArcGIS, inverse distance weighting, unit price maps for each entire state were developed as Fig. 86.2. The maps were color-coded with increasing unit prices when colors changed from blue to red. The maps can visually show that Montana had higher unit prices of the work item than Iowa in 2013.

The above statement was further checked by statistical analysis. The average unit price of the projects in Iowa was $643.11 per ton, while the average in Montana was $688.44 per ton. Since some tests of normality (i.e., normal Q-Q plot, Kolmogorov-Smirnov, and Shapiro-Wilk) proved that two samples of unit prices were not normally distributed, a non-parametric test was used to determine whether two sets of data are significantly different. The Mann-Whitney U test rejected the hypothesis that the two distributions are the same with the significance level of 0.05, which was consistent with the results from the maps.

86.4 Conclusion

This study presents a GIS-based visual analytics framework for highway projects by applying technologies and techniques from web data extraction, information extraction, databases, semantic similarity, and GIS. Depending on the availability of data and scale of research, part of or entire framework can be utilized for visual analytics. The framework allows for not only typical data analyses of un-spatial data (e.g., frequencies, average values, and mean differences) but also spatial analysis (e.g., variations of unit prices over an area). Applying the entire framework can enable a large-scale study of highway projects for several DOTs, a region, or even the whole nation. Our goals are to evaluate project performances regarding cost and productivity, quantify the effects of location and time on the performances, and visualize the results through interactive maps.

The case study in this paper is just limited in comparison of a work item from two DOTs using the inverse distance weighting interpolation method. In future research, it can be expanded to a regional level such as for Midwest region including twelve states. Major bid items in highway projects, which account for the majority of the total cost, will be identified and then matched across DOTs for comparisons. Different kinds of interpolation methods will be utilized and evaluated to find the one that can produce unit price maps with smallest errors. The variances of unit prices across states will

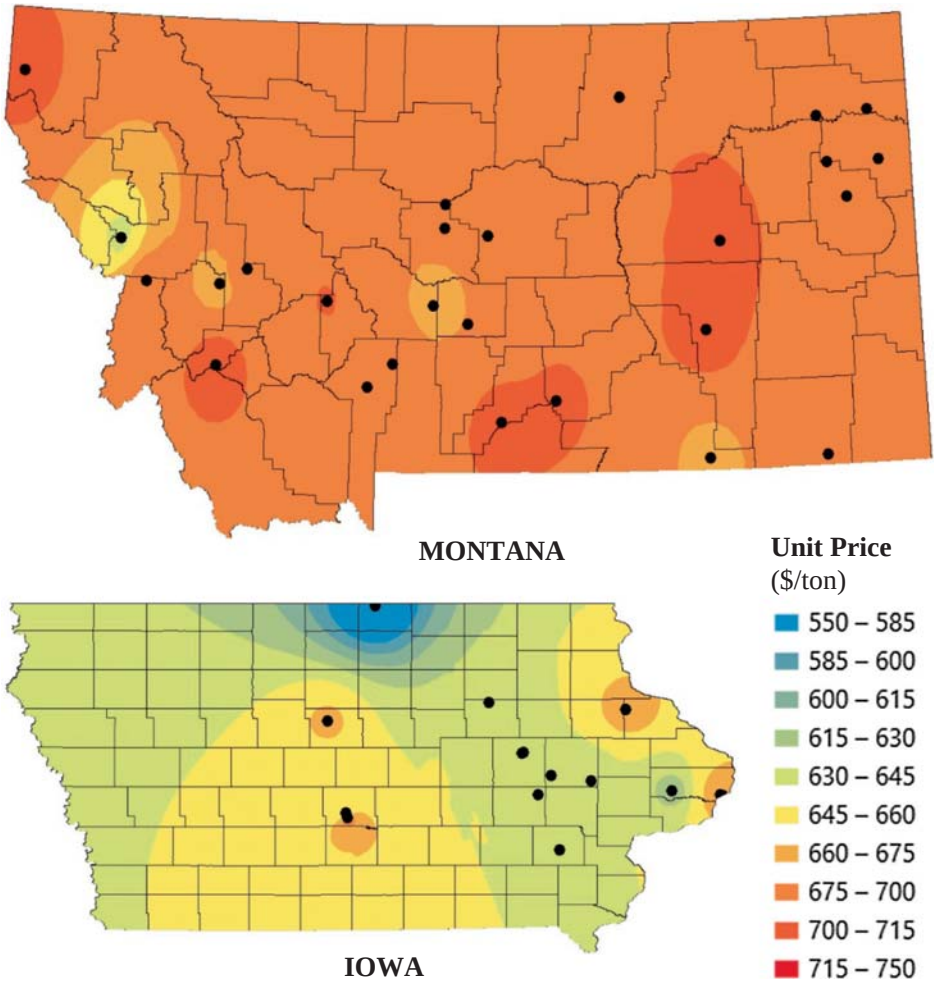

Fig. 86.2 Unit price maps of asphalt binder/cement PG 64-28, 2013

be visualized to support decision-making processes such as unit price estimations. Further interviews with practitioners from DOTs could be implemented to explain any differences between DOTs.

References

1. Aggarwal, C., Zhai, C.: Mining Text Data. Springer Science & Business Media (2012)
2. Burrough, P.A., McDonnell, R.A., Lloyd, C.D.: Principles of Geographical Information Systems. Oxford University Press (2015)
3. Cambria, E., White, B.: Jumping NLP curves: a review of natural language processing research [review article]. IEEE Comput. Intell. Mag. **9** (2), 48–57 (2014). https://doi.org/10.1109/MCI.2014.2307227
4. Chia-Hui, C., Kayed, M., Girgis, M.R., Shaalan, K.F.: A survey of web information extraction systems. IEEE Trans. Knowl. Data Eng. **18** (10), 1411–1428 (2006). https://doi.org/10.1109/TKDE.2006.152
5. Childs, C.: Interpolating surfaces in ArcGIS spatial analyst. ArcUser, July-September, 3235, 569 (2004)
6. Clementini, E., Di Felice, P.: A comparison of methods for representing topological relationships. Inf. Sci. Appli. **3**(3), 149–178 (1995)
7. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. GATE: an architecture for development of robust HLT applications. Paper presented at the Proceedings of the 40th annual meeting on association for computational linguistics, Philadelphia, Pennsylvania, 2002
8. Din, A., Hoesli, M., Bender, A.: Environmental variables and real estate prices. Urban stud. **38**(11), 1989–2000 (2001)
9. Eberly, S., Swall, J., Holland, D., Cox, B., Baldridge, E.: Developing spatially interpolated surfaces and estimating uncertainty. U. S. Environ. Prot. Agency (2004)
10. Erk, K.: Vector space models of word meaning and phrase meaning: a survey. Lang. Linguist. Compass **6**(10), 635–653 (2012). https://doi.org/10.1002/lnco.362

11. Ferrara, E., De Meo, P., Fiumara, G., Baumgartner, R.: Web data extraction, applications and techniques: a survey. Knowl.-Based Syst. **70**, 301–323 (2014). https://doi.org/10.1016/j.knosys.2014.07.007
12. Gogar, T., Hubacek, O., Sedivy, J.: Deep Neural Networks for Web Page Information Extraction. Cham (2016)
13. Grefenstette, E., Dinu, G., Zhang, Y.-Z., Sadrzadeh, M., Baroni, M.: Multi-step regression learning for compositional distributional semantics. arXiv preprint arXiv:1301.6939 (2013)
14. Ha, H., Olson, J.R., Bian, L., Rogerson, P.A.: Analysis of heavy metal sources in soil using kriging interpolation on principal components. Environ. Sci. Technol. **48**(9), 4999–5007 (2014)
15. Harispe, S., Ranwez, S., Janaqi, S., Montmain, J.: Semantic similarity from natural language and ontology analysis. Synth. Lect. Hum. Lang. Technol. **8**(1), 1–254 (2015). https://doi.org/10.2200/S00639ED1V01Y201504HLT027
16. Harris, Z.S.: Distributional structure. Word, 10 (2–3): 146–162. Reprinted in Fodor, J.A, Katz, J.J. (eds.), Readings in the Philosophy of Language. Prentice-Hall, Englewood Cliffs, NJ (1954)
17. Ian, H., Sarah, C., Steve, C.: An Introduction to Geographical Information Systems. Pearson Education India (2010)
18. Iorio, A.D., Peroni, S., Poggi, F., Vitali, F., Shotton, D.: Recognising document components in XML-based academic articles. Paper presented at the Proceedings of the 2013 ACM symposium on document engineering, Florence, Italy, 2013
19. Jiang, J.: Information extraction from text. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data, pp. 11–41. Springer and Boston, MA, US (2012)
20. Johnston, K., Ver Hoef, J.M., Krivoruchko, K., Lucas, N.: Using ArcGIS geostatistical analyst (Vol. 380): Esri Redlands (2001)
21. Laender, A.H.F., Ribeiro-Neto, B.A., da Silva, A.S., Teixeira, J.S.: A brief survey of web data extraction tools. SIGMOD Rec. **31**(2), 84–93 (2002). https://doi.org/10.1145/565117.565137
22. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. Paper presented at the international conference on machine learning, 2014
23. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
24. Mooney, R.: Relational learning of pattern-match rules for information extraction. Paper presented at the Proceedings of the sixteenth national conference on artificial intelligence, 1999
25. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. Paper presented at the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014
26. Piskorski, J., Yangarber, R.: Information extraction: past, present and future. In: Poibeau, T., Saggion, H., Piskorski, J., Yangarber, R. (eds.) Multi-source, Multilingual Information Extraction and Summarization, pp. 23–49. Springer, Berlin, Heidelberg (2013)
27. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. Paper presented at the Proceedings of the 2003 Conference of the North American chapter of the association for computational linguistics on human language technology—Volume 1, Edmonton, Canada, 2003
28. Turian, J., Ratinov, L., Bengio, Y.: Word representations: a simple and general method for semi-supervised learning. Paper presented at the Proceedings of the 48th annual meeting of the association for computational linguistics, Uppsala, Sweden, 2010
29. Webster, J.J., Kit, C.: Tokenization as the initial phase in NLP. Paper presented at the Proceedings of the 14th conference on computational linguistics—Volume 4, Nantes, France, 1992
30. Xiao, Y., Gu, X., Yin, S., Shao, J., Cui, Y., Zhang, Q., Niu, Y.: Geostatistical interpolation model selection based on ArcGIS and spatio-temporal variability analysis of groundwater level in piedmont plains, northwest China. SpringerPlus **5**(1), 425 (2016)
31. Zhai, Y., Liu, B.: Web data extraction based on partial tree alignment. Paper presented at the Proceedings of the 14th international conference on World Wide Web, Chiba, Japan, 2005
32. Zhao, H., Kit, C.: Integrating unsupervised and supervised word segmentation: the role of goodness measures. Inf. Sci. **181**(1), 163–183 (2011). https://doi.org/10.1016/j.ins.2010.09.008