Design of a Small and Medium Enterprise Growth Prediction Model Based on Web Mining

Yiea-Funk Te and Irena Pletikosa Cvijikj

ETH Zurich, Weinberstrasse 56/58, 8092 Zurich, Switzerland
{fte,ipletikosa}@ethz.ch

Abstract. Small and medium enterprises (SMEs) play an important role in the economy of many countries. Still, due to the highly turbulent business environment, SMEs experience more severe challenges in maintaining and expanding their business. To support SMEs at improving their competitiveness, researchers recently turned their focus on applying web mining (WM) to build growth prediction models. WM enables automatic and large-scale collection and analysis of potentially valuable data from various online platforms, thus bearing a great potential for extracting SME growth factors, and enhancing existing SME growth prediction models. This study aims at developing an automated system to collect business-relevant data from the Web and predict future growth trends of SMEs by means of WM and machine learning (ML) techniques.

Keywords: Data Mining, Web Mining, SME Growth Prediction.

1 Research Problem

Small and medium enterprises (SMEs) play an important role in the world economy. SMEs represent 95% of all businesses, accounting for 66% of the total employment and 55% of the total production [11]. However, existing studies show that the current business environment is highly turbulent and influenced by modern information and communication technologies, globalization and employee mobility [1,14]. Additionally, the growing number of SMEs caused competition to become increasingly intensive, forcing SMEs to experience more severe challenges in maintaining their existence and expanding their business. Thus, understanding the SMEs success factors is of great importance.

In order to support SMEs at improving their competitiveness, scholars and practitioners have been analyzing growth and success factors of SMEs for decades. With the emergence of big data, the focus was placed on applying data mining techniques to build novel risk and growth prediction models. However, existing models only include few data types, e.g. financial or operational data, and thus cannot fully explain the complex context of SME growth [13]. Moreover, data is mostly obtained via questionnaires, which is a time-consuming process, or is provided by financial institutes, thus not publicly available and sensitive to privacy issues. In addition, existing models mostly focus on risk evaluation and bankruptcy prediction. Although numerous studies on SME growth prediction exist, studies applying data mining techniques are scarce.

Recently, web mining (WM) has emerged as a new approach towards obtaining valuable business insights. WM enables an automated and large scale collection and analysis of potentially valuable data from the web such as the national commercial register and company websites. While WM methods have been frequently studied to anticipate growth of sales volume for e-commerce platforms, their application for assessment of SME growth factors is still scarce. Considering the large and increasing amount of data freely available online, WM bears a great potential in revealing valuable information hidden in web, which can be further used to build a SME growth prediction model.

2 Research Objectives

I aim at developing a growth prediction model for SMEs based on publicly available data. Therefore, I explore the potential of using WM to access a wide range of factors influencing growth from various web data sources. Finally, I aim at developing an automated system to collect business-relevant data from the web and predict future growth trends of SMEs by means of WM and data mining techniques. The envisioned system should serve as an "early recognition system" for future growth opportunities.

Business growth can be measured as (1) non-financial, i.e., growth of employment, customer satisfaction and loyalty, or (2) financial, such as growth of revenues, profits and assets. The focus of my work is on the revenue growth, due to its importance to the economy. To address the mentioned issues, I state the following research questions:

- RQ1. How can web data be used to leverage SME growth modelling? - The proposed research project explores the potential of WM to access growth factors from the web in order to build a SME growth prediction model. In order to answer this generally formulated question, three sub-questions are stated as follows.
- RQ2. Which growth factors can be extracted from the web? - Various web data sources have been identified which will be further analyzed with regard to the usability and accuracy to build a SME growth prediction model. Given a large set of well-studied growth factors, the goal is to identify suitable web data sources for feature extraction through WM.
- RQ3. From all the potential growth factors, which one prove to be discriminative? - Since a large variety of features are included in the growth prediction model, it is crucial to understand which factors are essential. To answer this question, different machine learning (ML) algorithms capable of feature selection or feature ranking will be applied [19]. The goal is to (1) confirm existing SME growth studies in a data-driven and model-based manner, (2) identify new growth correlates to extend our understanding of SME growth mechanisms, and finally (3) remove redundant features, thus reduce overfitting and improve the generalization of the models.
- RQ4. Which ML techniques can improve current state-of-the-art SME growth models? - Different ML algorithms will be tested with the goal to optimize the performance in predicting the future growth development of SMEs. Furthermore, in order to evaluate the additional value of our web data based growth model, the accuracy of our final model will be compared to a baseline growth model built on conventional, i.e. not publicly accessible data provided by a large insurance company.

3 State of the Art

3.1 SME Growth Research

The literature on business growth models dates back to 1967 and has proliferated since then into different streams addressing specific industries and business sizes. Lippitt and Schmidt [8] developed a general growth model for all sizes of businesses by examining how personality development theories influences the creation, growth and maturation of a business. Steinmetz [17] qualitatively analyzed the growth of small enterprises by partitioning the growth curve of small enterprises into different stages and assessing the characteristic attributes of each stage. A qualitative study conducted by Scott and Bruce [16] suggested a model for small business growth supporting managers to plan for future growth. The proposed model isolates five growth stages characterized by an unique combination of firm attributes. As a small company goes through different growth stages, attributes such as the management style, organizational structure and the use of technology changes. Although stage models are widely accepted among researchers and practitioners, stage models are criticized on some counts [12]. In particular, empirical research are only conducted on small sample sizes and specific types of businesses via questionnaires studies and thus, stage models are not generalizable.

3.2 Data Mining in SME Risk and Growth Research

Data mining techniques such as artificial neural network and decision tree have been extensively studied with a strong focus on SME risk evaluation and bankruptcy prediction rather than SME growth modelling. An early study indicated that backpropagation neural network (BPNN) were the most popular ML techniques among researchers in the finance and business domain during the 1990's [20]. For instance, Zhang et al. [22] provide a comprehensive review of Artificial Neural Network (ANN) applications for bankruptcy prediction. Their findings indicated that ANN perform significantly better than logistic regression models. West [19] investigated the credit scoring accuracy of various ANN models (e.g. multilayer perceptron, mixture-of-experts, radial basis function) in comparison with traditional methods such as logistic regression and discriminant analysis model. Consistent with the findings of Zhang et al. [22], ANN models perform slightly better than traditional methods.

Other techniques widely applied in the domain of risk evaluation and bankruptcy prediction includes decision trees (DT) and their ensemble variations such as random forests (RF). For instance, Fantazzini and Figini [4] developed a model based on RF for SME credit risk measurement and compared its performance with the traditional logistic regression approach. They came to the conclusion that both models provided similar results in terms of performance, highlighting the potential of RF for credit risk modelling. Another recent and more application-oriented study conducted by Ozgulbas and Koyuncugil [13] proposed an early warning system based on DT-algorithms for SMEs to detect risk profiles. The proposed system uses financial data to identify risk indicators and early warning signs, and create risk profiles for the classification of SMEs into different risk levels.

In summary, data mining techniques such as ANN and DT are extensively used for SME risk evaluation and bankruptcy prediction. However, research reporting data mining based SME growth prediction cannot be identified. Furthermore, current prediction models usually include one type of data sources and thus cannot explain the whole and complex context of SME growth [13].

3.3 Web Mining in E-commerce and SME Growth Research

The web is a popular and interactive medium with intense amount of freely available data. It is a collection of documents, audios, videos and other multimedia data [9]. With billions of web pages available in the web, it is a rapidly growing key source of information, presenting an opportunity for businesses and researchers to derive useful knowledge out of it. Therefore, WM research for knowledge discovery has emerged.

WM has been proved very useful in the business world, especially in e-commerce [6,7,10,15,18] . For instance, Lin and Ho [7] proposed a system, which extracts content information from a set of web pages with a goal of extracting the informative content. Morinaga et al. [10] presented a system for finding the reputation of products from the internet to support marketing and customer relationship management in order to increase the sales volume. Finally, Thorleuchter et al. [18] analyzed the impact of textual information from e-commerce companies' websites on their commercial success by extracting web content data from the most successful top 500 worldwide companies.

While WM methods has been well researched and used in the field of e-commerce research to increase the sales volume, it has barely been applied for SME growth research. The study conducted by Antlová, Popelínsky and Tandler [2] demonstrated the potential of WM for SME growth prediction. In their paper, they examined the relationship between long-term growth of SMEs and a web presentation. Li et al. [6] recently explored micro-level characteristics and impacts of external relationships such as government or university relations on the growth of SMEs. However, these studies only focus on the information available in company websites and thus, restricting the amount and spectrum of growth factors. Hence, further research exploiting the full potential of web data for SME growth prediction is required.

4 Methodology

4.1 Research Design

In order to answer the stated research questions, the research is structured as follows. In the first step, a large set of well-studied growth factors is identified from an extensive literature review, which will serve as a groundwork for feature generation.

In the preliminary study, a SME growth model based on conventional and mostly well-structured data will be developed, serving as a baseline model. Therefore, data provided by a large Swiss insurance company will be used. The data consist of a wide range of firm attributes. Furthermore, the data contain information about the annual revenues, which will serve as "growth label" for supervised ML (Figure 1: left).

In the main study, WM methods for automatic data collection and pre-processing will be applied. Various web sources, such as company websites, commercial register websites and social media will be studied. The retrieved web data will be used to develop a web data based growth prediction model. Finally, the accuracy of the final model will be compared to that of the baseline model (Figure 1: right).

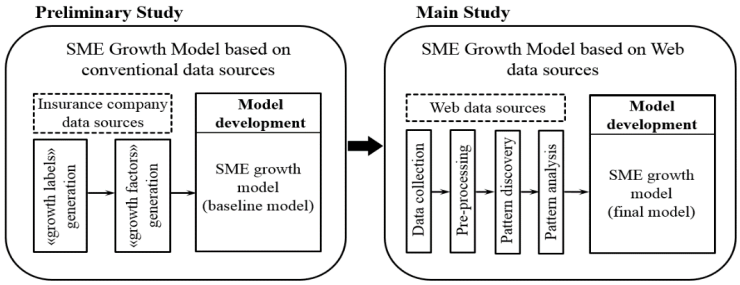

Fig. 1. Research design.

4.2 Identification of Growth Factors

The current business environment is influenced by a variety of firm internal and external factors [21], as illustrated in Figure 2. Firm-internal factors can be divided into: (1) the characteristics of the firm, such as firm attributes (age, size, location) and firm strategies (marketing, training strategies), and (2) the characteristics of the entrepreneur, such as socio-demographic characteristics (age, gender, family background) and the personality of the entrepreneur (need for achievement, risk-taking propensity). Firm-external factors can be divided into factors reflecting (1) the immediate and (2) the contextual environment. The immediate environment includes supplier and customer relationship, competition, labor market and resource market. In contrast, the contextual environment comprises macro-environmental factors, such as economic, political, socio-cultural, technological and legal influences on the growth of businesses, which can emanate from local and national sources, but also from international developments [21].

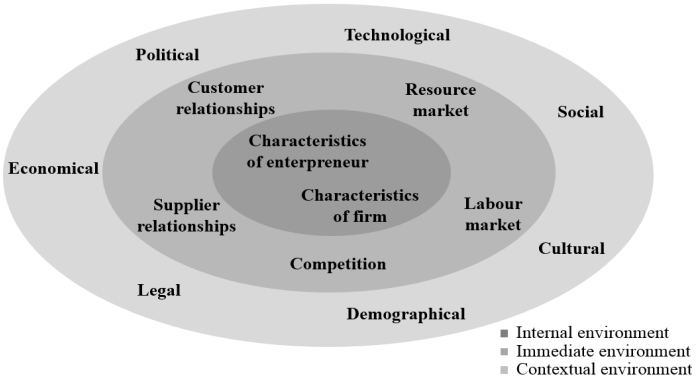

Fig. 2. Factors influencing growth of SMEs.

4.3 Data Collection

Two different types of data will be used: (1) not publicly available data provided by a large Swiss insurer, and (2) publicly available web data.

The data provided by the insurance company cover a 7-year period from 2010-2016 and contain detailed information about a large set of Swiss SMEs including their annual revenue. This data source will be used to: (1) build the baseline growth model, and (2) serve as ground truth data (i.e. growth-labeled data) to train the final model.

The web data related to a large set of Swiss SMEs with known revenues (i.e. ground truth) are collected. Therefore, the usability of various web data sources will be evaluated with respect to the identified growth factors. The data sources include governmental websites such as the Swiss Federal Statistical Office, commercial websites such as the Central Business Names Index, social media like Twitter and LinkedIn, and company websites. First, I plan to manually inspect the data sources in order to assess their usability. The next step is to develop a python-based program which allows an automated extraction of previously identified growth factors for a large set of companies. Table 1 summarizes the data sources which will be examined.

Table 1. Web data sources to be examined.

Web sources	Description
Central Business Names Index	Federal registry about firm location, founding year, board members, capital structure, legal authority
Swiss Federal Statistical Office	Granular information about the Swiss socio-demographics, economics, culture, education and political orientation
Federal Institute of Intellectual Property	Intellectual protected property
Job postings	Indicator for demand, skill level and sourcing
Twitter, LinkedIn, Xing	Information about company owners, company postings and mentions
Company websites	Information about the firm such as location, team, products, partners and company news

5 Expected Outcome

In my PhD thesis, I aim to enable the prediction of SME growth based on publicly available web data, currently not a common approach in the domain of SME growth prediction. Therefore, various web data sources will be analyzed with respect to their usability for SME growth prediction. WM will applied to collect web data and feature generation, whereas different ML algorithms will be studied to build a SME growth model with high accuracy. Furthermore, my research aims to expand the existing

knowledge of several closely-related domains, including business intelligence, business information systems, web mining and SME growth research.

Furthermore, the research carried out in the scope of this thesis will have important practical implications. For the Swiss SME organizations, the insights generated in my thesis may support the Swiss SME organizations at understanding the success and failure of SMEs, thus strengthen their supportive role for SMEs. For the investment companies, the built system can be used to monitor the development of SMEs by mining the changes in the website of firms, serving as an "early recognition system" for future opportunities for growth. Finally, for SMEs, the system can be used to evaluate the characteristics of firms based on the information given in the web. The absence of important key success factors can be pointed out to firms, thus serving as a consulting program.

6 Stage of the Research

At the time of writing, the factors influencing growth of SMEs have been determined through an extensive literature review. The potential web sources for feature generation are identified and will be further investigated with regard to the usability to build a SME growth model. In particular, the data from the Central Business Names Index and Swiss Federal Statistical Office are completely collected, which gives us an overall picture of the SME landscape and the general socio-economic situation in Switzerland. Furthermore, the data from company websites of our ground truth (approximately 9000 SMEs) are collected by a web crawler and stored in the HTML format. In the next step, the focus lies in the extraction of firm characteristics such as age of firm, size, financial resources and human capital by means of text mining. A first SME growth model will be built based on the collected data. Therefore, different ML algorithms will be used und compared with regard to the accuracy. In an iterative and incremental approach, features reflecting the characteristics of the entrepreneur and environmental factors will be extracted from other web sources and added to the growth model.

7 Advice Sought

I hope to get valuable feedback from senior WM researchers on potential techniques suitable for extraction of growth factors from the web data sources. In particular, handling company websites is very challenging because the information available on company websites is not standardized, and varies according to the company and how it wishes to present itself. In addition, I would like to discuss the prediction ML algorithms I plan to use, as well as and further useful datasets and potential collaborations.

References

1. Antlová, K.: Motivation and Barriers of ICT Adoption in Small and Medium-Sized Enterprises. E+M Ekonomie a Management 22(2), 140-154 (2009).

2. Antlová, K., Popelínsky, L., Tandler, L.: Long term growth of SME from the view of ICT competencies and web presentations. E+ M Ekonomie a management (4), 125 (2011).
3. Breiman, L.: Decision-tree forests. Machine Learning 45(1), 5-32 (2001).
4. Fantazzini, D., Figini, S.: Random survival forests models for SME credit risk measurement. Methodology and Computing in Applied Probability 11(1), 29-45 (2009).
5. Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27(2), 83-85 (2005).
6. Li, Y., Arora, S., Youtie, J., Shapira, P.: Using web mining to explore Triple Helix influences on growth in small and mid-size firms. Technovation (2016).
7. Lin, S. H., Ho, J. M.: Discovering informative content blocks from Web documents. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 588-593. ACM (2002).
8. Lippitt G. L., Schmidt, W. H.: Crises in a developing organization. Harvard Business Review (1967).
9. Malarvizhi, R., Saraswathi, K.: Web Content Mining Techniques Tools & Algorithms-A Comprehensive Study. International Journal of Computer Trends and Technology (IJCTT) 4(8) (2013).
10. Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T.: Mining product reputations on the web. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 341-349. ACM (2002).
11. OECD. Small and Medium-Sized Enterprises In Turkey Issues And Policies. Organization For Economic Cooperation And Development. OECD Press (2004).
12. O'Farrell, P. N., Hitchens, D. M.: Alternative theories of small-firm growth: a critical review. Environment and Planning A 20(10), 1365-1383 (1988).
13. Ozgulbas, N., Koyuncugil, A. S.: Risk Classification of SMEs by Early Warning Model Based on Data Mining. World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering 6(10), 2649-2660 (2012).
14. Post, H. A.: Building a Strategy on Competencies. Long range Planning 30(5), 733–740 (1997).
15. Saini, S., Pandey, H. M.: Review on Web Content Mining Techniques. International Journal of Computer Applications 118(18) (2015).
16. Scott, M., Bruce, R.: Five Stages of Growth in Small Business. Long Range Planning, pp. 45-52 (1987).
17. Steinmetz, L. L.: Critical stages of small business growth: When they occur and how to survive them. Business horizons 12(1), 29-36 (1969).
18. Thorleuchter, D., Van Den Poel, D.: Predicting e-commerce company success by mining the text of its publicly-accessible website. Expert Systems with Applications 39(17), 13026-13034 (2012).
19. West, D.: Neural network credit scoring models. Computers & Operations Research 27(11), 1131-1152 (2000).
20. Wong, B. K., Lai, V. S., Lam, J.: A bibliography of neural network business applications research: 1994–1998. Computers & Operations Research 27, 1045–1076 (2000).
21. Worthington, C. & Britton, I.: The business environment. Financial Times, Harlow (2006).
22. Zhang, G., Hu, M. Y., Patuwo, B. E., Indro, D. C.: Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. European journal of operational research 116(1), 16-32 (1999).