Analyzing and Visualizing Relationship Networks

Xiaoyi Lu

Supervisor Name : Wei Liu
School of Computer Science and Software Engineering

Client Mentor Name : Sam Winterton, Alan Thompson
CEED Client: Corporate Governance Risk

Abstract

This project is designed to detect the hidden relationships between risks, model them and visualize them. The technique is based on a mixture of text clustering and text classification. Firstly, the raw risks are preprocessed and transformed into the desired format. Then Latent Dirichlet Allocation (LDA) is applied to the risks as a cluster technique to discover the hidden thematic topics that can be used to group risks. After obtaining the topic structure, new risks can be classified into these topics using techniques like Naïve Bayes (NB). The approach is able to not only model the relationships between existing risks, but also model the relationships between "new" risks and "old" risks. In addition, a JavaScript library called D3 has been utilized to visualize the relationship network.

1. Introduction

Risk management is a process of identifying risks, assessing risks and taking actions to minimize the adverse impacts of risks. Risks are core concepts in the process of risk management, which need to be modeled and stored in a database. As the number of risks in the database grows, it becomes more difficult to organize them.

Basically, the project focuses on the relationships between risks. A risk is potentially linked to another risk if the two risks have similar contents. For instance, "Inadequate levels of accommodation" should be linked to "Insufficient construction accommodation" and "Adverse Weather Event" should be associated with "Unanticipated weather events".

The project aims to develop a tool suite, which helps a company to allocate their limited resources in a more efficient manner. It is known that significant resources must be put in place to keep track of a risk. Similar risks mean redundant risks in some sense. By identifying similar risks, time and resources can be saved. In addition, by identifying the matrix of all similar risks, the most influential or centroid risks can be identified, which indicates that more weight should be placed on them. The tool suite will be able to group similar risks together so that they can be managed easily.

1.1 The state of the art

Identifying relationships requires a mixture of text classification (Sebastiani 2002) and text clustering. Classification techniques like NB (Rennie et al. 2003) and clustering techniques like LDA (Blei, Ng & Jordan 2003) have been successfully applied in many applications.

They have been used to identify patients at risk for developing cancer from free text radiology reports (Garla, Taylor, & Brandt 2013). To date no one has applied these techniques to solve problems in the risk management industry. The goal of the project is to bring natural language processing techniques into the domain of risk management and deal with real-life issues.

1.2 Objectives

The aim of the project is to develop a tool suite which can be used to identify and analyze the hidden relationships between risks. The tool suite should be able to:

1. Predict potential links between risks based on their contents.
2. Users should be able to navigate through similar risks.
3. Similar risks should be organized into same groups so that they can be managed easily.
4. New risks should be able to fit into the existing groups.
5. Provide a way of overviewing how risks are distributed.

2. Methodology

The project consists of three major stages: preprocessing the raw data, clustering the existing risks and classifying new risks into the clusters.

2.1 Preprocessing

Raw risk data are processed and transformed into a suitable representation for later tasks. Basically, risks are treated as "bags" of words. Each risk is then represented as a vector. Each entry of the vector corresponds to a word in the risk. The value in the entry is the frequency count of the corresponding word. In order to improve efficiency and effectiveness, stop words removal and word stemming are performed. In addition, infrequent words are also erased to minimize the size of vector space.

2.2 Clustering

In the clustering stage, LDA is used to cluster the existing risks. LDA is a simplest topic model, which is a set of algorithms for discovering the main themes that pervade a large collection of data (Steyvers & Griffiths 2007).
A topic is defined as a distribution over a fixed vocabulary (Blei, Ng & Jordan 2003). For example, the "weather" topic has words about weather with high probability, like cyclone, lightning. The "tug" topic would has words about tug with high probability, like port, harbor.

The foundation behind LDA is that risks exhibit multiple topics. For example, 30% of the risk might be talking about marine construction and 70% might be talking about environmental protection. Note that "marine construction" and "environmental construction" are topics that are summarized by experts. LDA could only output topics as distributions over words.

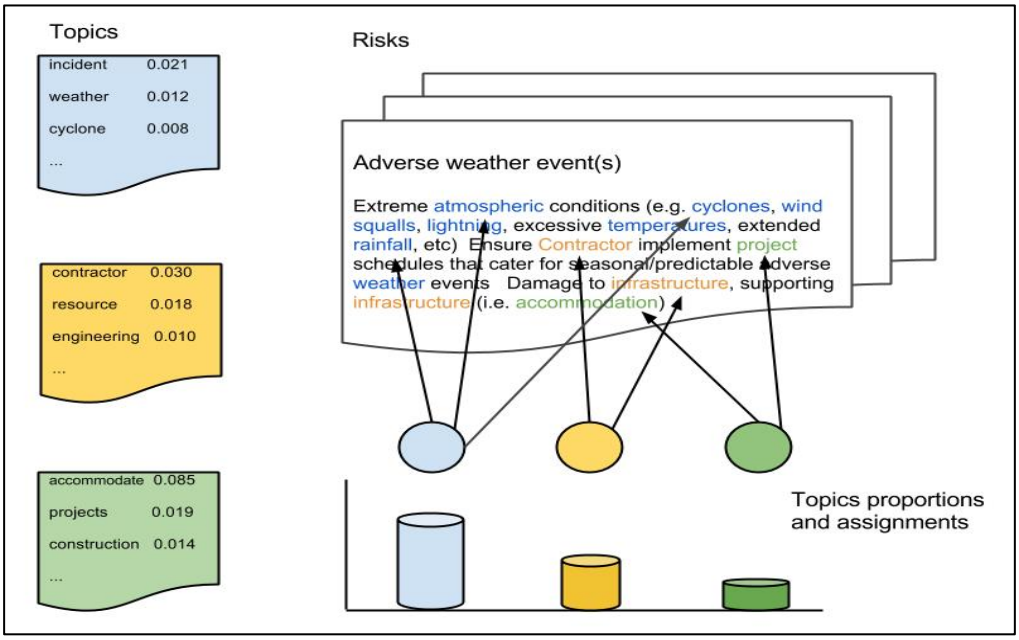

Figure 1 Topic proportions and assignments

Figure 1 shows that each risk is a mixture of topics. Each word in a risk is drawn from one of the topics. By inferring the underlying topics, the risks are naturally linked together through these topics. In addition, the topic structure provides a way to organize these risks.

2.3 Classification

The tool suite is not only able to divide the existing risks into groups, it is also capable of situating new risks into the existing topic structure. The classification can be viewed as a way to model the relationships between "future" risks and existing risks. In other words, the existing topics are predefined categories. New risks are then classified and labeled by the topics. It should be a multi-label classification task since each risk exhibits a mixture of topics. But to make things simpler, we only label the risk as the topic that has the highest probability. Take the example of the risk we discussed before, the risk should be labeled as "environmental protection" rather than "marine construction".

Currently, three different classification techniques are utilized to classify new risks: NB, K nearest neighbors (kNN) and Logistic Regression (LR). In order to obtain a better predictive performance, majority voting is used as an ensemble method to combine all the three classifiers. Basically, the ensemble would choose the topic that receives the largest votes as the label of the risk. For instance, given a new risk, NB classifier chooses topic 1, kNN classifier chooses topic 3 and LR classifier chooses topic 1, then the ensemble would choose topic 1 as the label of the risk. Majority Voting is expected to outperform any of the individual classifier.

3. Results and Discussion

The first part of the project is to find out the thematic topics that group the risks. Below are topics extracted from a real client database, presented using D3 JavaScript library.

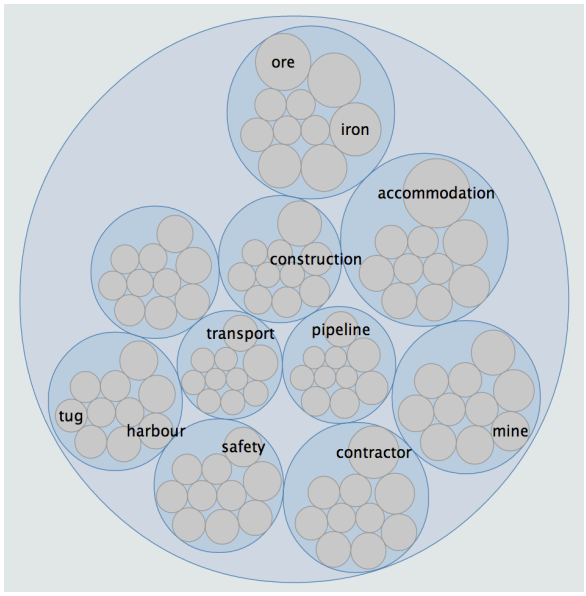

Figure 2 Topics extracted from a real client db

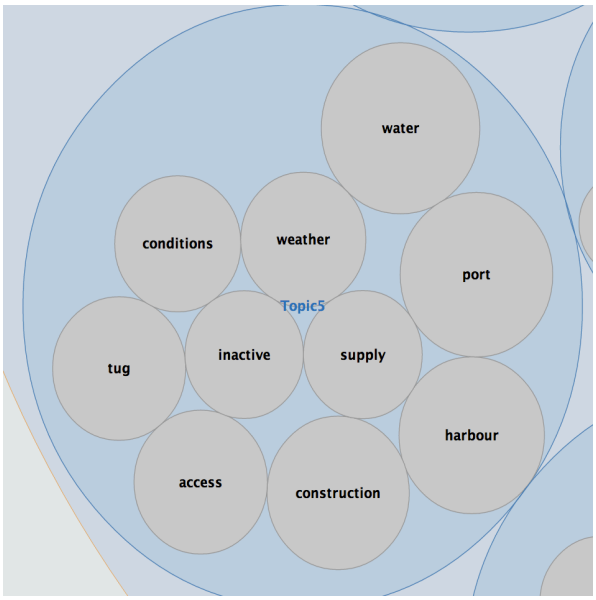

Figure 3 Top ranked words in Topic 5

Figure 2 shows ten topics extracted from a real client database, represented by ten circles inside the big one. For each topic circle, there are ten little sub-circles that are shown in Figure 3, which represent the top ten ranked words in each topic. It is obvious from figure 3 that the circles containing "water" and "port" are larger than others. This is because these words are more likely to show up in this topic. Actually, "water" appears 308 times in topic 5 while "port" shows up 285 times in the same topic. It can be inferred from the word list that topic 5 is talking about water issues as well as port/harbor construction issues.

In this case, relationships between risks are represented by the topics and the words. They provide two different levels of linking the risks. For example, if the user clicks the circle containing the word "weather", then all the risks in topic 5 that are associated with weather would be returned to the user. This is offering a "zoom-in" view of the topic to the user. Then the user can zoom out the topic and have an overview of all the risks under the topic.

In addition, relationships between risks can also be interpreted as the similarities between risks. For example, if the similarity between two risks is higher than the threshold defined beforehand, the two risks are potentially associated.

Figure 4 is a similarity matrix showing the similarities of a set of risks. All color-coded cells indicate potential linkages between risks on the left and risks at the top. Deeper color implies the closer the two risks are. Similarly, figure 5 displays the potential relationship network in a different angle. Every edge in the graph links two risks indicating they are close to each other.

The second part of the project is to be able to situate the new risks into the existing topic structure. The classification results are presented as a bar chart, shown in Figure 6 below. all three classifiers have performed quite well in classifying risks into the right topics. Among those, NB stands out with the highest accuracy and precision, followed by LR and kNN.

Majority Voting performs almost the same as the NB. But it goes against our original expectation that ensemble methods should outperform any of the individual methods. This is because only three classifiers are used. If both of LR and kNN make the wrong predictions, then the performance of the Majority Voting will be affected adversely.

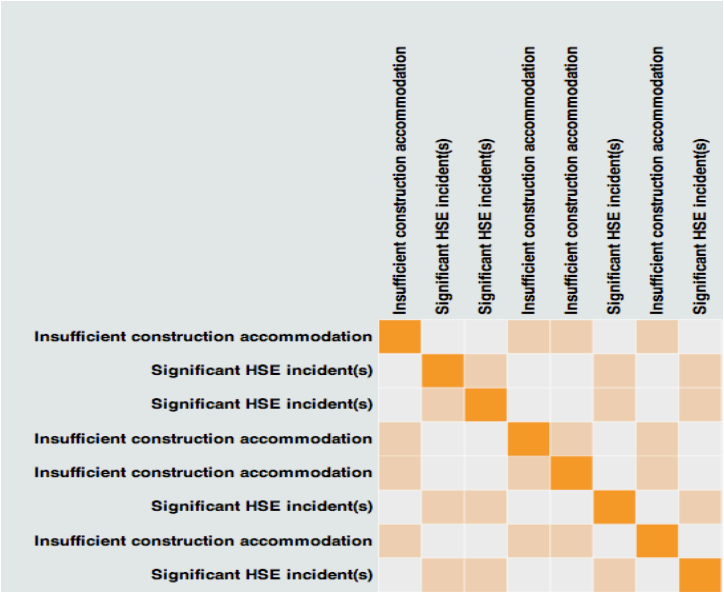

Figure 4 Similarity matrix of risks

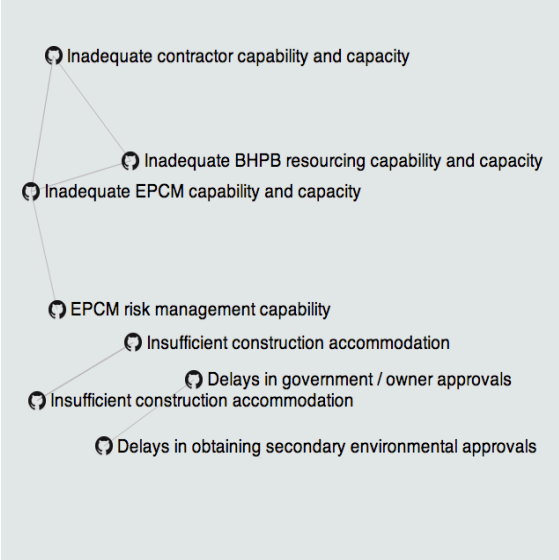

Figure 5 Force-directed graph of linked risks

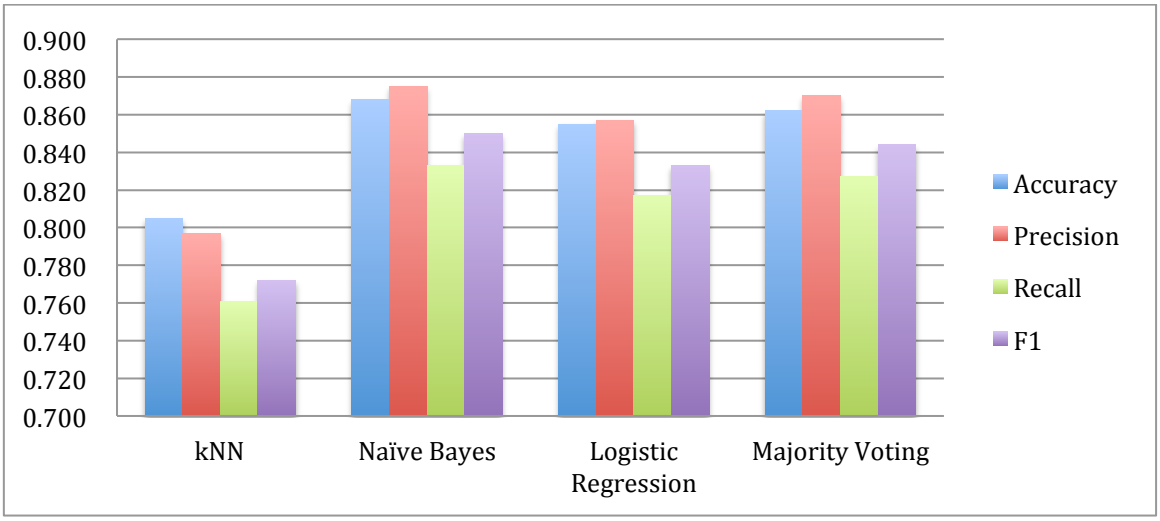

Figure 6 Classification Performance Comparison

4. Conclusions and Future Work

The tool suite is able to model the relationships between risks through topics and similarities.

Topic modeling is an emerging field in machine learning, and there are many exciting new directions for research.

1. The current topic model is only one level deep. By applying the LDA in the extracted topics, hierarchical topic structures would be expected, which provide a more fine-grained approach to group data.

2. For now, only three classifiers are used to classify risks. By adding more classifiers, ensemble method is expected to outperform any of the individual methods.

5. References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003) Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3** pp. 993-1022.

Garla, V., Taylor, C., & Brandt, C. (2013) Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. *Journal of biomedical informatics*.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003) Tackling the poor assumptions of naive bayes text classifiers. *In Proceedings of the Twentieth International Conference on Machine Learning*, **3** pp. 616-623.

Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, **34** (1) pp. 1-47.

Steyvers, M., & Griffiths, T. (2007) Probabilistic topic models. *Handbook of latent semantic analysis*, **427**(7) pp.424-440.