Hunting for the Black Swan: Risk Mining from Text

Jochen L. Leidner and **Frank Schilder**
Thomson Reuters Corporation
Research & Development
610 Opperman Drive, St. Paul, MN 55123 USA
FirstName.LastName@ThomsonReuters.com

Abstract

In the business world, analyzing and dealing with risk permeates all decisions and actions. However, to date, *risk identification*, the first step in the risk management cycle, has always been a manual activity with little to no intelligent software tool support. In addition, although companies are required to list risks to their business in their annual SEC filings in the USA, these descriptions are often very high-level and vague.

In this paper, we introduce *Risk Mining*, which is the task of identifying a set of risks pertaining to a business area or entity. We argue that by combining Web mining and Information Extraction (IE) techniques, risks can be detected automatically before they materialize, thus providing valuable business intelligence.

We describe a system that induces a risk taxonomy with concrete risks (e.g., interest rate changes) at its leaves and more abstract risks (e.g., financial risks) closer to its root node. The taxonomy is induced via a bootstrapping algorithms starting with a few seeds. The risk taxonomy is used by the system as input to a risk monitor that matches risk mentions in financial documents to the abstract risk types, thus bridging a lexical gap. Our system is able to automatically generate company specific "risk maps", which we demonstrate for a corpus of earnings report conference calls.

1 Introduction

Any given human activity with a particular intended outcome is bound to face a non-zero likelihood of failure. In business, companies are exposed to market risks such as new competitors, disruptive technologies, change in customer attitudes, or a changes in government legislation that can dramatically affect their profitability or threaten their business model or mode of operation. Therefore, any tool to assist in the elicitation of otherwise unforeseen risk factors carries tremendous potential value.

However, it is very hard to identify risks exhaustively, and some types (commonly referred to as the *unknown unknowns*) are especially elusive: if a *known unknown* is the established knowledge that important risk factors are known, but it is unclear whether and when they become realized, then an *unknown unknown* is the lack of awareness, in practice or in principle, of circumstances that may impact the outcome of a project, for example. Nassim Nicholas Taleb calls these "black swans" (Taleb, 2007).

Companies in the US are required to disclose a list of potential risks in their annual Form 10-K SEC fillings in order to warn (potential) investors, and risks are frequently the topic of conference phone calls about a company's earnings. These risks are often reported in general terms, in particular, because it is quite difficult to pinpoint the *unknown unknown*, i.e. what kind of risk is concretely going to materialize. On the other hand, there is a stream of valuable evidence available on the Web, such as news messages, blog entries, and analysts' reports talking about companies' performance and products. Financial analysts and risk officers in large companies have not enjoyed any text analytics support so far, and risk lists devised using questionnaires or interviews are unlikely to be exhaustive due to small sample size, a gap which we aim to address in this paper.

To this end, we propose to use a combination of Web Mining (WM) and Information Eextraction (IE) to assist humans interested in risk (with respect to an organization) and to bridge the gap between the general language and concrete risks. We describe our system, which is divided in two main parts: (a) an offline **Risk Miner** that facilitates the *risk identification* step of the risk management process, and an online (b) **Risk Monitor** that supports the *risk monitoring* step (cf. Figure 2). In addition, a **Risk Mapper** can aggregate and visualize the evidence in the form of a *risk map*. Our risk mining algorithm combines Riloff hyponym patterns with recursive Web pattern bootstrapping and a graph representation.

We do not know of any other implemented end-to-end system for computer-assisted risk identification/visualization using text mining technology.

Proceedings of the ACL 2010 System Demonstrations, pages 54–59,
Uppsala, Sweden, 13 July 2010.

2 Related Work

Financial IE. IE systems have been applied to the financial domain on Message Understanding Contest (MUC) like tasks, ranging from named entity tagging to slot filling in templates (Costantino, 1992).

Automatic Knowledge Acquisition. (Hearst, 1992) pioneered the pattern-based extraction of hyponyms from corpora, which laid the groundwork for subsequent work, and which included extraction of knowledge from to the Web (e.g. (Etzioni et al., 2004)). To improve precision was the mission of (Kozareva et al., 2008), which was designed to extract hyponymy, but they did so at the expense of recall, using longer *dual anchored patterns* and a pattern linkage graph. However, their method is by its very nature unable to deal with low-frequency items, and their system does not contain a chunker, so only single term items can be extracted. De Saenger et al. (De Saeger et al., 2008) describe an approach that extracts instances of the "trouble" or "obstacle" relations from the Web in the form of pairs of fillers for these binary relations. Their approach, which is described for the Japanese language, uses support vector machine learning and relies on a Japanese syntactic parser, which permits them to process negation. In contrast, and unlike their method, we pursue a more general, open-ended search process, which does not impose as much a priori knowledge. Also, they create a set of pairs, whereas our approach creates a taxonomy tree as output. Most importantly though, our approach is not driven by frequency, and was instead designed to work especially with rare occurrences in mind to permit "black swan"-type risk discovery.

Correlation of Volatility and Text. (Kogan et al., 2009) study the correlation between share price volatility, a proxy for risk, and a set of trigger words occurring in 60,000 SEC 10-K filings from 1995-2006. Since the disclosure of a company's risks is mandatory by law, SEC reports provide a rich source. Their trigger words are selected a priori by humans; in contrast, risk mining as exercised in this paper aims to find risk-indicative words and phrases automatically.

Kogan and colleagues attempt to find a regression model using very simple unigram features based on whole documents that predicts volatility, whereas our goal is to automatically extract patterns to be used as alerts.

Speculative Language & NLP. Light et al. (Light et al., 2004) found that sub-string matching of 14 pre-defined string literals outperforms an SVM classifier using bag-of-words features in the task of speculative language detection in medical abstracts. (Goldberg et al., 2009) are concerned with automatic recognition of human wishes, as expressed in human notes for Year's Eve. They use a bi-partite graph-based approach, where one kind of node (content node) represents things people wish for ("`world peace`") and the other kind of node (template nodes) represent templates that extract them (e.g. "`I wish for ___`"). Wishes can be seen as positive Q, in our formalization.

3 Data

We apply the mined risk extraction patterns to a corpus of financial documents. The data originates from the StreetEvents database and was kindly provided to us by Starmine, a Thomson Reuters company. In particular, we are dealing with 170k earning calls transcripts, a text type that contains monologue (company executives reporting about their company's performance and general situation) as well as dialogue (in the form of questions and answers at the end of each conference call). Participants typically include select business analysts from investment banks, and the calls are published afterwards for the shareholders' benefits. Figure 1 shows some example excerpts. We randomly took a sample of N=6,185 transcripts to use them in our risk alerting experiments.[1]

4 Method

4.1 System

The overall system is divided into two core parts: (a) Risk Mining and (b) Risk Monitoring (cf. Figure 2). For demonstration purposes, we add a (c) Risk Mapper, a visualization component. We describe how a variety of risks can be identified given a normally very high-level description of risks, as one can find in earnings reports, other financial news, or the risk section of 10-K SEC filings. Starting with rather abstract descriptions such as *operational risks* and hyponym-inducing pattern "`< RISK > such as *`", we use the Web to mine pages from which we can harvest additional,

[1] We could also use this data for risk mining, but did not try this due to the small size of the dataset compared to the Web.

CEO: As announced last evening, during our third quarter, we will take the difficult but necessary step to seize [cease] manufacturing at our nearly 100 year old Pennsylvania House casegood plant in Lewisburg, Pennsylvania as well as the nearby Pennsylvania House dining room chair assembly facility in White Deer. Also, the three Lewisburg area warehouses will be consolidated as we assess the logistical needs of the casegood group's existing warehouse operations at an appropriate time in the future to minimize any disruption of service to our customers. This will result in the loss of 425 jobs or approximately 15% of the casegood group's current employee base.
Analyst: Okay, so your comments – and I guess I don't know – I can figure out, as you correctly helped me through, what dollar contribution at GE. I don't know the net equipment sales number last quarter and this quarter. But it sounded like from your comments that if you exclude these fees, that equipment sales were probably flattish. Is that fair to say? **CEO:** We're not breaking out the origination fee from the equipment fee, but I think in total, I would say flattish to slightly up.

Figure 1: Example sentences from the earnings conference call dataset. Top: main part. Bottom: Q&A.

and eventually more concrete, candidates, and relate them to risk types via a transitive chain of binary IS-A relations. Contrary to the related work, we use a base NP chunker and download the full pages returned by the search engine rather than search snippets in order to be able to extract risk phrases rather than just terms, which reduces contextual ambiguity and thus increases overall precision. The taxonomy learning method described in the following subsection determines a risk taxonomy and new risks patterns.

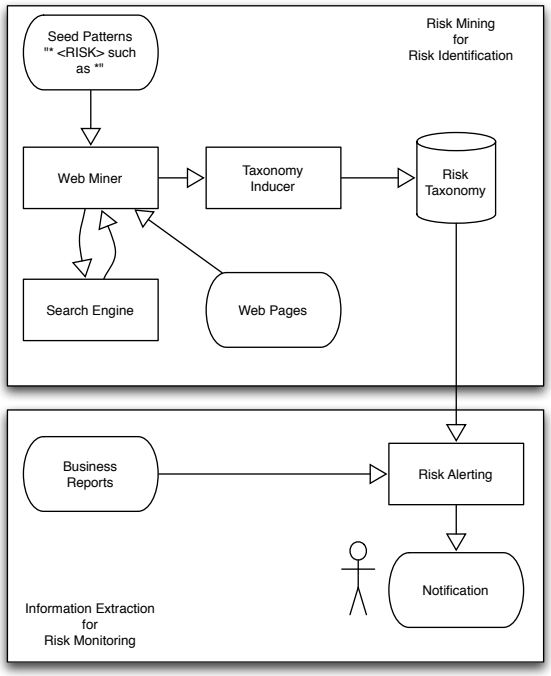

Figure 2: The risk mining and monitoring system architecture

The second part of the system, the Risk Monitor, takes the risks from the risk taxonomy and uses them for monitoring financial text streams such as news, SEC filings, or (in our use case) earnings reports. Using this, an analyst is then able to identify concrete risks in news messages and link them to the high-level risk descriptions. He or she may want to identify operational risks such as fraud for a particular company, for instance. The risk taxonomy can also derive further risks in this category (e.g., faulty components, brakes) for exploration and drill-down analysis. Thus, news reports about faulty breaks in (e.g. Toyota) or volcano outbreaks (e.g. Iceland) can be directly linked to the risk as stated in earnings reports or security filings.

Our Risk Miner and Risk Monitor are implemented in Perl, with the graph processing of the taxonomy implemented in SWI-Prolog, whereas the Risk Mapper exists in two versions, a static image generator for R[2] and, alternatively, an interactive Web page (DHTML, JavaScript, and using Google's Chart API). We use the Yahoo Web search API.

4.2 Taxonomy induction method

Using frequency to compute confidence in a pattern does not work for risk mining, however, because mention of particular risks might be rare. Instead of frequency based indicators (n-grams, frequency weights), we rely on two types of structural confidence validation, namely (a) previously identified risks and (b) previously acquired structural patterns. Note, however, that we can still use PageRank, a popularity-based graph algorithm, because multiple patterns might be connected to a risk term or phrase, even in the absence of frequency counts for each (i.e., we interpret popularity as having multiple sources of support).

1. Risk Candidate Extraction Step. The first step is used to extract a list of risks based on high precision patterns. However, it has been shown that the use of such patterns (e.g., *such as*) quickly lead to an decrease in precision. Ideally, we want to retrieve specific risks by re-applying the the extract risk descriptions:

[2] `http://www.r-project.org`

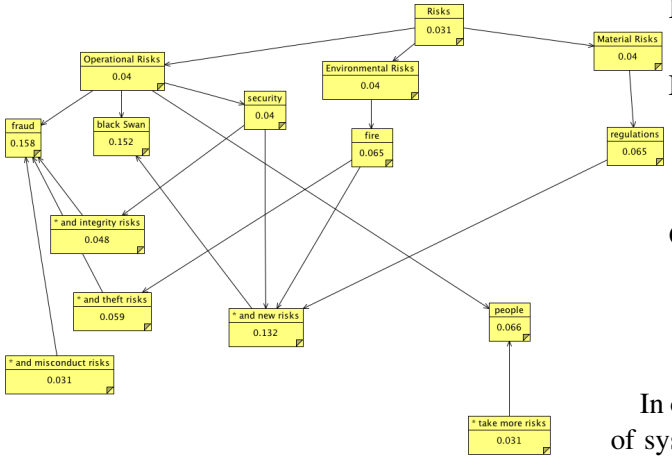

Figure 3: A sample IS-A and Pattern network with sample PageRank scores.

(a) Take a seed, instantiate "< SEED > such as *" pattern with seed, extract candidates:

Input: risks

Method: apply pattern "< SEED > such as < INSTANCE > ", where < SEED > = risks

Output: list of instances (e.g., *faulty components*)

(b) For each candidate from the list of instances, we find a set of additional candidate hyponyms.

Input: faulty components

Method: apply pattern "< SEED > such as < INSTANCE > ", where < SEED > = faulty components

Output: list of instances (e.g., *brake*)

2. Risk Validation. Since the Risk Candidate extraction step will also find many false positives, we need to factor in information that validates that the extracted risk is indeed a risk. We do this by constructing a possible pattern containing this new risk.

(a) Append "* risks" to the output of 1(b) in order to make sure that the candidate occurs in a risk context.

Input: brake(s)

Pattern: "brake(s) * risk(s)"

Output: a list of patterns (e.g., *minimize such risks*, *raising the risk*)

(b) extract new risk pattern by substituting the risk candidate with < RISK > ; creating a limited number of variations

Input: list of all patterns mined from step 2 (a)

Method: create more pattern variations, such as "< RISK > minimize such risks", "raising the risk of < RISK > " etc.

Output: list of new potential risks (e.g., *deflation*), but also many false positives (e.g., *way*, as in *The best way to minimize such risks*).

In order to benefit from any human observations of system errors in future runs, we also extended the system so as to read in a partial list of predefined risks at startup time, which can guide the risk miner; while technically different from active learning, this approach was somewhat inspired by it (but our feedback is more loose).

3. Constructing Risk Graph. We have now reached the point where we constructed a graph with risks and patterns. Risks are connected via IS-A links; risks and patterns are connected via PATTERN links. Note that there are links from risks to patterns and from patterns to risks; some risks back-pointed by a pattern may actually not be a risk (e.g., *people*). However, this node is also not connected to a more abstract risk node and will therefore have a low PageRank score. Risks that are connected to patterns that have a high authority (i.e., pointing to by many other links) are highly ranked within PageRank (Figure 3). The risk *black Swan*, for example, has only one pattern it occurs in, but this pattern can be filled by many other risks (e.g., fire, regulations). Hence, the PageRank score of the black swan is high similar to well known risks, such as *fraud*.

4.3 Risk alerting method

We compile the risk taxonomy into a trie automaton, and create a second trie for company names from the meta-data of our corpus. The Risk Monitor reads the two tries and uses the first to detect mentions of risks in the earning reports and the second one to tag company names, both using case-insensitive matching for better recall. Optionally, we can use Porter stemming during trie construction and matching to trade precision for even higher recall, but in the experiments reported here this is not used. Once a signal term or phrase matches, we look up its risk type in a hash table, take a note of the company that the current earnings report is about, and increase the frequency

```
liquidity IS-A financial risks
credit IS-A financial risks
direct risks IS-A financial risks
fraud IS-A financial risks
irregular activity IS-A operational risks
process failure IS-A operational risks
human error IS-A operational risks
labor strikes IS-A operational risks
customer acceptance IS-A IT market risks
interest rate changes IS-A capital market risks
uncertainty IS-A market risks
volatility IS-A mean reverting market risks
copyright infringement IS-A legal risks
negligence IS-A other legal risks
an unfair dismissal IS-A the legal risks
Sarbanes IS-A legal risks
government changes IS-A global political risks
crime IS-A Social and political risks
state intervention IS-A political risks
terrorist acts IS-A geopolitical risks
earthquakes IS-A natural disaster risks
floods IS-A natural disaster risks
global climate change IS-A environmental risks
severe and extreme weather IS-A environmental risks
internal cracking IS-A any technological risks
GM technologies IS-A tech risks
scalability issues IS-A technology risks
viruses IS-A the technical risks
```

Figure 4: Selected financial risk tuples after Web validation.

count for this ⟨company; risk type⟩ tuple, which we use for graphic rendering purposes.

4.4 Risk mapping method

To demonstrate the method presented here, we created a visualization that displays a *risk map*, which is a two dimensional table showing companies and the types of risk they are facing, together with bubble sizes proportional to the number of alerts that the Risk Monitor could discover in the corpus. The second option also permits the user to explore the detected risk mentions per company and by risk type.

5 Results

From the Web mining process, we obtain a set of pairs (Figure 4), from which the taxonomy is constructed. In one run with only 12 seeds (just the risk type names with variants), we obtained a taxonomy with 280 validated leave nodes that are connected transitively to the risks root node.

Our resulting system produces visualizations we call "risk maps", because they graphically present the extracted risk types in aggregated form. A set of risk types can be selected for presentation as well as a set of companies of interest. A risk map display is then generated using either R (Figure 5) or an interactive Web page, depending on the user's preference.

Qualitative error analysis. We inspected the output of the risk miner and observed the following classes of issues: (a) chunker errors: if phrasal boundaries are placed at the wrong position, the taxonomy will include wrong relations. For example, deictic determiners such as "this" were a problem (e.g. `that IS-A indirect risks`) before we introduced a stop word filter that discards candidate tuples that contain no content words. Another prominent example is "short term" instead of the correct "short term risk"; (b) *semantic drift*[3]: due to polysemy, words and phrases can denote risk and non-risk meanings, depending on context. Talking about risks even a specific pattern such as "such as" [sic] is used by authors to induce a variety of perspectives on the topic of risk, and after several iterations negative effects of type (a) errors compound; (c) off-topic relations: the seeds are designed to induce a taxonomy specific to risk types. As a side effect, many (correct or incorrect) irrelevant relations are learned, e.g. `credit and debit cards is-a money transfer`. We currently discard these by virtue of ignoring all relations not transitively connected with the root node `risks`, so no formalized domain knowledge is required; (d) overlap: the concept space is divided up differently by different writers, both on the Web and in the risk management literature, and this is reflected by multiple category membership of many risks (e.g. is cash flow primarily an operational risk or a financial risk?). Currently, we do not deal with this phenomenon automatically; (e) redundant relations: at the time of writing, we do not cache all already extracted and validated risks/non-risks. This means there is room for improvement w.r.t. runtime, because we make more Web queries than strictly necessary. While we have not evaluated this system yet, we found by in-

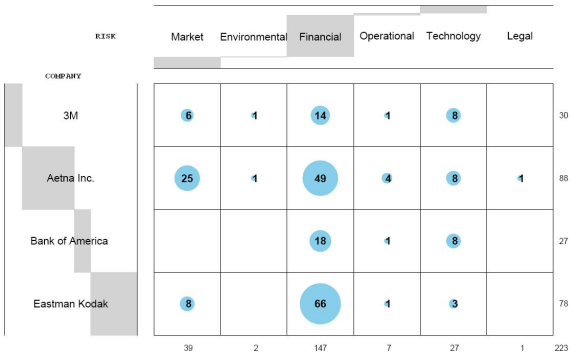

Figure 5: An Example Risk Map.

[3] to use a term coined by Andy Lauriston

specting the output that our method is particularly effective for learning natural disasters and medical conditions, probably because they are well-covered by news sites and biomedical abstracts on the Web. We also found that some classes contain more noise than others, for example operational risk was less precise than financial risk, probably due to the lesser specificity of the former risk type.

6 Summary, Conclusions & Future Work

Summary of Contributions.

In this paper, we introduced the task of risk mining, which produces patterns that are useful in another task, risk alerting. Both tasks provide computational assistance to risk-related decision making in the financial sector. We described a special-purpose algorithm for inducing a risk taxonomy offline, which can then be used online to analyze earning reports in order to signal risks. In doing so, we have addressed two research questions of general relevance, namely how to extract *rare patterns*, for which frequency-based methods fail, and how to use the Web to bridge the *vocabulary gap*, i.e. how to match up terms and phrases in financial news prose with the more abstract language typically used in talking about risk in general.

We have described an implemented demonstrator system comprising an offline risk taxonomy miner, an online risk alerter and a visualization component that creates visual risk maps by company and risk type, which we have applied to a corpus of earnings call transcripts.

Future Work. Extracted negative and also positive risks can be used in many applications, ranging from *e-mail alerts* to determinating *credit ratings*. Our preliminary work on risk maps can be put on a more theoretical footing (Hunter, 2000). After studying further how output of risk alerting correlates[4] with non-textual signals like share price, risk detection signals could inform human or trading decisions.

Acknowledgments. We are grateful to Khalid Al-Kofahi, Peter Jackson and James Powell for supporting this work. Thanks to George Bonne, Ryan Roser, and Craig D'Alessio at Starmine, a Thomson Reuters company, for sharing the StreetEvents dataset with us, and to David Rosenblatt for discussions and to Jack Conrad for feedback on this paper.

[4]Our hypothesis is that risk patterns can outperform bag of words (Kogan et al., 2009).

References

Marco Costantino. 1992. Financial information extraction using pre-defined and user-definable templates in the LOLITA system. *Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING 1992), vol. 4*, pages 241–255.

Stijn De Saeger, Kentaro Torisawa, and Jun'ichi Kazama. 2008. Looking for trouble. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, pages 185–192, Morristown, NJ, USA. Association for Computational Linguistics.

Oren Etzioni, Michael J. Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. 2004. Web-scale information extraction in KnowItAll: preliminary results. In Stuart I. Feldman, Mike Uretsky, Marc Najork, and Craig E. Wills, editors, *Proceedings of the 13th international conference on World Wide Web (WWW 2004), New York, NY, USA, May 17-20, 2004*, pages 100–110. ACM.

Andrew B. Goldberg, Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. 2009. May all your wishes come true: A study of wishes and how to recognize them. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 263–271, Boulder, Colorado, June. Association for Computational Linguistics.

Marti Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics (COLING 1992)*.

Anthony Hunter. 2000. Ramification analysis using causal mapping. *Data and Knowledge Engineering*, 32:200–227.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of the Joint International Conference on Human Language Technology and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic class learning from the web with hyponym pattern linkage graphs. In *Proceedings of ACL-HLT*, pages 1048–1056, Columbus, OH, USA. Association for Computational Linguistics.

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *BioLINK 2004: Linking Biological Literature, Ontologies and Databases*, pages 17–24. ACL.

Nassim Nicholas Taleb. 2007. *The Black Swan: The Impact of the Highly Improbable*. Random House.