The NASD Securities Observation, News Analysis & Regulation System (SONAR)

Henry Goldberg, Dale Kirkland, Dennis Lee, Ping Shyr, Dipak Thakker *

NASD (National Association of Securities Dealers), 9513 Key West Avenue, Rockville, MD 20850
[GoldberH, KirklanD, LeeDE, ShyrP, ThakkerD]@NASD.com

Abstract

The Securities Observation, News Analysis, and Regulation (SONAR) system was developed by NASD to monitor the Nasdaq, Over the Counter (OTC), and Nasdaq-Liffe (futures) stock markets for potential insider trading and fraud through misrepresentation. SONAR has been in operational use at NASD since December 2001, processing approximately 10,000 news wires stories and SEC filings, evaluating price/volume models for 25,000 securities, and generating 50-60 alerts (or "breaks") per day for review by several groups of regulatory analysts and investigators. In addition, SONAR has greatly expanded surveillance coverage to new areas of the market and increased accuracy significantly over an earlier break detection system. SONAR makes use of several AI and statistical techniques, including NLP text mining, statistical regression, rule-based inference, uncertainty, and fuzzy matching. Sonar combines these enabling technologies in a system designed to deliver a steady stream of high-quality breaks to the analysts for further investigation. Additional components including visualization, text search agents, and flexible displays add to the system's utility. Finally, SONAR is designed as a knowledge-based system. Domain knowledge is maintained by knowledge engineers working closely with the regulatory analysts. In this way, SONAR is adaptable to new market conditions, data sources, and regulatory concerns.

Introduction

NASD is a self-regulatory organization that oversees and regulates several securities markets, the largest one being the Nasdaq Stock Market. Other regulated markets include the OTC Bulletin Board, Third Market, Pink Sheet Market, and the Nasdaq Liffe Market for single stock futures.

NASD conducts surveillance for potential violative market activity. When such activity is discovered, NASD will conduct an investigation and, when merited, (1) take disciplinary action on broker-dealers and/or individuals registered with NASD or (2) refer the investigation results to another regulatory (e.g., the US SEC) or law enforcement (e.g., the US Department of Justice) body for action. Two such activities are (1) Insider Trading (IT) on material non-public information and (2) Fraud against investors involving misrepresentation, usually in text, as to the true nature of a publicly traded company.

When we use the term "insider trading" in this paper, the context is insider trading on information that an investor would want to know (i.e., it is material) but that has not yet been disseminated to the public.

Fraud can occur in a variety of forms. This application addresses fraud against investors that involves misrepresentation of the nature of a publicly traded company by persons who should know otherwise. There are other types of fraudulent activity that do not represent legitimate interests but are designed to mislead (e.g., reporting a trade late to mask front running a large order, wash sales, money laundering). Other systems at NASD address these.

Business Units: The business department with primary responsibility for the surveillance of stock markets regulated by the NASD is the Market Regulation Department (MRD). The MRD conducts the reviews, investigations, disciplinary actions, and referrals as the situation demands. Insider Trading and Fraud are investigated by two teams (within the Surveillance and Compliance Section) of up to a dozen analysts each. The analysts typically have a background in finance, the securities industry, or law.

History of NASD Automated Surveillance and Knowledge Based Systems: The MRD has conducted automated surveillance for many years. Beginning in 1988, the MRD developed a system called Stock Watch Automated Tracking (SWAT) that covered insider trading and fraud. The replacement to SWAT is the Securities Observation News Analysis & Regulation (SONAR) system that is described in this paper.

Beginning in 1996 the MRD developed the Advanced Detection System (ADS). This system was designed to find patterns and practices of violative behavior of rules by firms (securities brokers). [Kirkland 1999] ADS is the initial knowledge-based system developed by the MRD. Along with deploying it, we developed a knowledge management process, including a board consisting of department executives to manage how we make use of ADS. [Senator 2000] It is within that framework that we developed the SONAR system as a knowledge-based automated surveillance system

Task Description

The following sections describe the tasks performed by analysts in the Insider Trading and Fraud teams and the challenge of gathering, analyzing, and associating massive amounts of text and market data as evidence for both

*The authors of this paper are employees of the NASD. The views expressed herein are those of the authors and do not represent an official policy statement of NASD.

detection and investigation. SONAR is aimed at automating as many of these tasks as is feasible and integrating smoothly with the analysts' work flow.

Insider Trading: Each trading day the Insider Trading Team should review market activity in every issue (i.e. stock) that has material news released by an issuer (i.e. company) that day. The analyst attempts to determine whether the market reacted to the material news, and whether it is likely that insider trading occurred. This amounts to reading the current news, reading the news in the recent past for the issuer, and reviewing what has occurred with trade prices and cumulative volume.

We know from past cases that insider trading is highly correlated with the type of news that is released. Based upon several decades of experience with insider trading cases, we know that certain kinds of stories are more likely to be associated with that event. Over 85% of past Insider Trading cases were based on material news of five types – Product announcements, Earnings announcements, Regulatory approvals or denials, Mergers and acquisitions, or Research reports (which we call PERM-R events). In addition, the trending of price and volume prior to and after news are considered.[1] If we suspect that insider trading could have occurred profitably, we start the process to identify those insiders and the recent trading activity they have undertaken. Also, we look for suspicious trading by persons who may have been tipped about the inside information by insiders.

Fraud: Each trading day the Fraud Team should review market activity in every issue for potential fraud by misrepresentation. An issue may or may not have material news released by the issuer on a given day. This does not change the fact that a check should be done. The analyst should read the news for any evidence that could be related to a pump-and-dump scheme[2]. That analyst should look for potential evidence of touting (i.e., the pump before the dump) that may occur on the issuer website, in spam e-mail messages, or on popular securities chat rooms. Furthermore, the analyst should check the recent Edgar filings to establish, among other things, the financial assets of the issuer, the company officers and directors, and the nature of the business. If the combined evidence gathered points to the likelihood of a pump-and-dump scheme, the process begins of contacting broker-dealers, company officers, and others to gather supporting or mitigating evidence. When the evidence is conclusive, the Fraud Team takes the necessary action to stop the activity and, if necessary, remove the perpetrators from the business.

Evidence Gathering

Data Sources: Prompt and accurate response to possible market violations depends on obtaining and analyzing reliable and up-to-date evidence to support investigations. These are collected from many sources, including market data, news (particularly financial news), SEC corporate filings, spam e-mail, websites, chat rooms, and many of NASD's internal documents such as SEC referrals, complaint data, and disciplinary history data.

One of the major goals of the SONAR system is to automate the evidence gathering, analyzing and linking process. Even after a potential concern is identified, an analyst still needs to review large amounts of market and text information to determine if there is an explanation for the apparent violation. Prior to SONAR, analysts reviewed information in raw formats (e.g. market data in tabular formats and full text of wire stories), from which it was difficult to discern relationships.

The primary data sources for SONAR (input and analyzed on a daily basis) consist of:

- Stories from four major news wire services: Dow Jones, Reuters, Bloomberg, and PR Newswire. On average about 8,500 – 10,000 news stories input each day from these sources although the number has been as high as 18,000 news stories in a day.
- The quarterly and annual filings collected by the SEC in its Edgar database provide corporate fundamentals, such as financial data, personnel data, business areas, and business development plans and locations. Each day there are about 1,000 Edgar filings of potential interest to the IT and Fraud teams.
- Market activity summarized from the daily transactions (trades, quotes, orders) of the subject markets. The average number of trades in all the markets of interest to the Insider Trading Team is about 5.5 million. Each day about 16,000 issues will have trading activity among the Nasdaq, OTCBB, and Pink Sheet markets.

No small group of analysts can handle this level of daily checking without substantial automated help from price and volume summarization and modeling, text mining and extraction, and evidence collection and linking.

Kinds of Evidence: News plays a pivotal role in the generation of breaks for many scenarios, and acts as a focal point for IT break detection. The results of text mining are news events, including news category, news nature, their associated issues, firms, and time frames.

[1] For example, insider trading to avoid a loss is more likely if the stock has been flat or trending down prior to bad news, since there is more to lose. SONAR employs several dozen trending scenarios as an heuristic to improve IT detection.

[2] While several different schemes are the subject of the Fraud team's surveillance, Pump-and-dump is the most critical to detect quickly because of the speed with which investor funds can be removed or laundered.

While analyzing news stories and Edgar filings, SONAR also collects evidence of suspiciousness. The system automatically groups the text-mined events into general categories as supporting information. Then flags are generated if the supporting events satisfy predetermined conditions and rules (e.g. payment in shares).

After gathering detailed market transactions, SONAR generates derived and aggregated attributes and builds issue profiles which can linked to text mined events. By doing this, the system not only provides analysts individual and detailed information, but also indicates the trends and relationships.

Existing cases and records of firms and individuals are collected in NASD internal documents by various departments. This valuable information provides violation or complaint histories for our reference. We are currently developing a general text input module for these internal materials.

Application Description

This section describes SONAR: how it works and what it is. After discussing the system concept and overall architecture, we will present the key functional modules of SONAR (see Figure 1):

- Text mining
- Post-extraction analysis (PEA)
- Market data analysis and modeling
- Post-detection analysis
- User interface
- Database of entities, relationships, and events

System Architecture

As a Break Detection System[3], SONAR provides comprehensive surveillance of securities markets with respect to particular regulatory concerns (insider trading and fraud). It is designed to be integrated into the organizational and data environment already in place at NASD. We recognized early on that the key activities which could be automated with AI technology were the detection, linking, and evaluation of evidence from primary sources relevant to daily market activity for a security. SONAR combines components to:

- detect evidence as it occurs in text sources (news wires and SEC filings)
- detect characteristic "events" in a space of price/volume-derived features of market activity, and
- combine this evidence in a meaningful way by assigning a probability-like score to each "security-day" which estimates the likelihood of several episodes of regulatory interest.

[3] See [Senator 2002] for a more general discussion of the roles and features of Break Detection Systems.

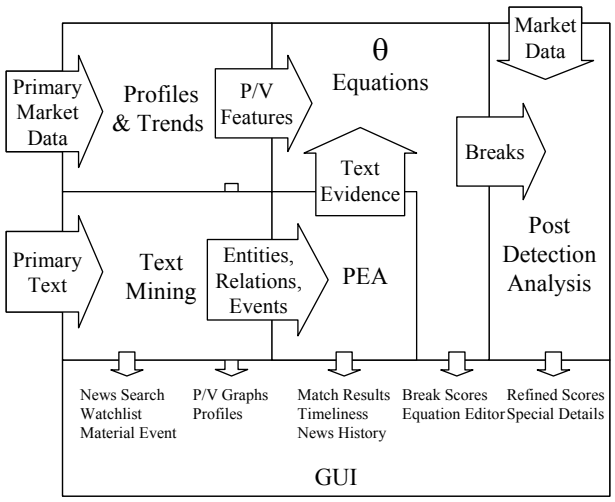

Figure 1: SONAR Conceptual Architecture

As pieces of evidence (e.g. a security symbol, a merger announcement, a fraud indicator) are detected, they are linked to a potential break for a particular security, trading day, and scenario. The statistical model which detects market events, also combines this evidence to produce an initial score. These scores may be adjusted based upon secondary source evidence in a post detection analysis stage. These actions are all performed by knowledge-based components which are described below. A web-based GUI integrates the results of these activities into a work product oriented break and evidence management system.

Text Mining

Within SONAR, natural language processing (NLP) software mines the daily stream of text documents from news wires Edgar for entities, relationships, and events relevant to IT and Fraud. Success of knowledge discovery of company news in SONAR is based on three criteria: pertinence, timeliness, and uniqueness. Is the extracted information pertinent to the subject matter of regulatory interest? When was a news story first reported? Is the extracted concept discovered multiple times? Once text mining has discovered key market events, the entities related to each market event are stored in a relational database. This architecture allows us to store, retrieve, and post-process key market events to support SONAR break detection and other market regulation investigation activities.

The text mining engine of SONAR is Clear Forest's text mining product – Clear Studio 3.0.[4] With support from Clear Forest the SONAR team developed a rule base supporting more than 90 top level predicates for both

[4] See: http://www.clearforest.com/

News Story and EDGAR filings. Rules extract key terms and concepts and are built using Clear Forest's proprietary DIAL Language. The building block of DIAL code objects allows for rule construction, local or global data types, flow control text scanning operators (global consumption, local consumption, cut operator), exception checking or constraints, and procedural calls to other C/C++ libraries and NLP libraries.

Consolidation and post-extraction analysis

As with any system dependant upon linking of evidence, consolidation of references to key entities is essential. [Goldberg 1995] Name-matching is extremely important to the SONAR system in identifying issuers and firms, and thus linking to the correct market activities. Our name matching process has two modes – dictionary lookup and fuzzy matching. SONAR creates a hashed ID[5] and stripped name for each new issue and company name based on predefined rules and add them to local and master dictionaries. In the fuzzy match process, the system tries to match first against the local dictionary when there exists a record within the same hash id. If there was no record matched in the local dictionary then we do a fuzzy match against the master dictionary, which relies on an heuristic distance measure. The high scoring matches are optionally associated with the text material. The name-matching algorithm can also handles aliases, nicknames, and abbreviations.

Following extraction of entities, relationships, and events from text, there are several analyses to improve upon the evidence stream. Post-Extraction Analysis (PEA) determines news category, uniqueness, timeliness, supporting information, and flags.

Different kinds of the announcements have different impacts in the market. For example, a new product announcement is usually not as significant as a merger/acquisition. Event categorization ranks and refines the events detected for an issue in a single day over (possibly) several stories.

Material news is often repeated, and the key to insider trading is to know when the material event is first publicized. Uniqueness evaluation performs a kind of topic tracking by comparing extracted events against previous ones in the database. Timeliness evaluation identifies and classifies the timeframes of the story from textual cues. By combining these two, we can avoid breaking on old or duplicated information.

Supporting information may be aggregated as flags, associated with an issue-day, for use by the Fraud break detection, if any one of several heuristics are satisfied.

[5] We currently use that most ancient of NLP algorithms, Soundex. This is an area for improvement.

Market data analysis and modeling

Measuring Market Activities: SONAR measures unusual price and volume movement in traded securities. A number of standard measures (called factors) are derived from the market data and are computed for all subject issues on a daily basis. (e.g. last sale price, rate of return over prior trading day, log volume traded) Some factors are derived by applying financial theories and models[6] to mimic what analyst routinely use. Others are derived using trending and scenario mapping.

We calculate price trends over a dynamic look-back period using a crossover method of two (short and long-term) exponential moving averages. The beginning of the period gives the system an indication of when the potential insider trading could have started. Once the look-back period is defined, it is used in measuring 1) the profit or loss avoidance potential of insider trading, and 2) pre-news price and volume trends and post-news market reaction. Along with the news nature (positive or negative) these are mapped to one of several scenarios (mentioned above) to provide evidence for refinement of the IT likelihood score.

Logistic Model for Detection and Combination of Evidence: The measures discussed above form a feature space for detectors of target activities. Points in this space are used in a logistic function to characterize a response variable, which presents a probability-like score of the target activity. Simply, the logit distribution is applied to a linear, weighted combination of features to produce an estimated probability between 0 and 1. The weights are trained using statistical regression over a selected training set. We have found that characteristics of the measurements of market activity support the use of logistic regression (LR) over multiple linear regression (MLR) because:

- Normal distribution does not adequately describe movement of stock price and their returns as well as log normal distribution does.
- MLR expects normal distribution of explanatory variables, whereas LR does not.
- MLR expects all the independent variables to be continuous.

By coding evidence from text mining numerically, as well as from the market measures space, we are able to use a single detector to combine both types of evidence and produce a probability-like score for each of several target activities. We call these detectors Theta Equations, after the response variable, and currently have about 15 in operation, covering various portions of the markets under surveillance. (e.g. IT Small Cap, Fraud Long Term Rise Pump and Dump, etc.)

[6] e.g. CAPM is used to correct for market return in issues determined to track the market.

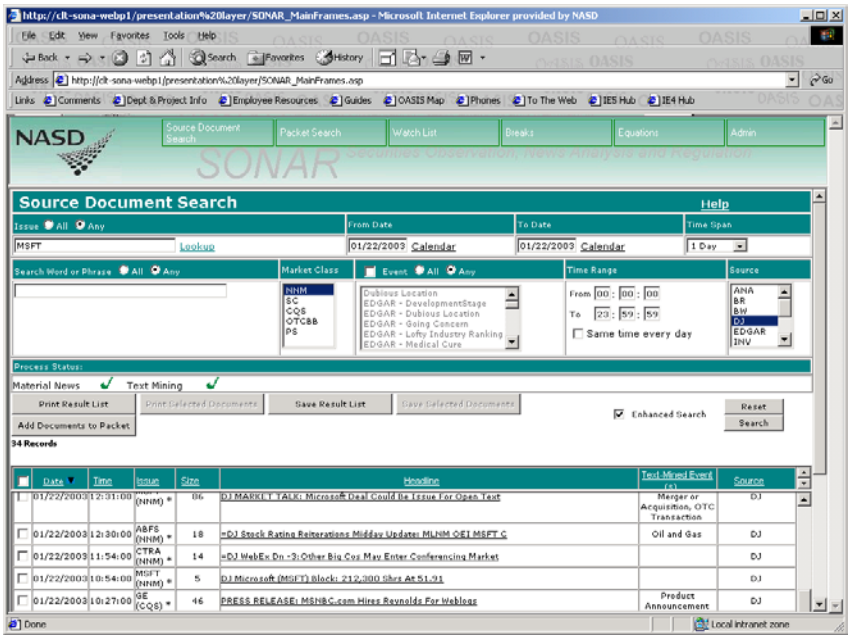

Figure 2: Source Document Search Screen

Post-detection analysis

Breaks are generated directly from the Theta Equations. Although the logistic regression models are efficient and effective, they have several shortcomings. It is hard to include all the business knowledge and data details we need to evaluate in these models. Also, the models cannot interact with one another across breaks. In order to further enhance the reliability of the break stream as a whole, a post-analysis has been included. A rule-based expert system[7] was added to provide a knowledge-based capability for post-detection analysis. As a first test of this module, we acquired and implemented rules pertaining to the inference of a type of insider trading known as Trading Ahead of Research Analyses. The CIA server inference is triggered by the initial detection of a stock research report in the news.

User Interface

SONAR has a web-based graphical user interface which consists of screens for document search, preparation and review of watch lists, search and review of breaks results, theta equation editing, and break administration. The user interface is designed to provide as much visibility as possible into the underlying data and decisions which SONAR has made.

Source document search is an important function. Users have a wide range of search parameters – words or phrases, document source, timeframe, market class, and text events. The source document search screen (see Figure 2) returns date, time, issue symbol, text headline, text-mined events and source as results. From the text headline, the system can bring up the full text for review. Both printing and saving capabilities were built in with this functionality for the user's convenience.

There are two major screens for logistic regression model development – theta equation display and factor editor. Knowledge specialists can edit the break score thresholds and factor coefficients in the theta equation display and the system will maintain an audit trail. The factor editor (see Figure 3) provides an editor to specify factor conditions and calculation. Users can provide factor descriptions, select which factor need to be included (e.g. look back period), apply business logic, and build control

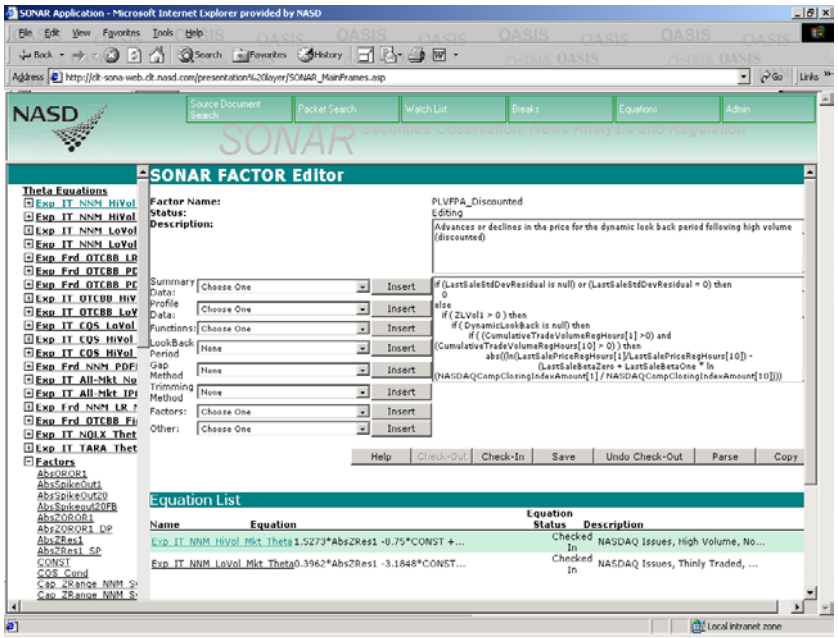

Figure 3: Factor Editor Screen

and conditions for a given factor.

Break details include tabular displays for viewing detailed issue information and trends associated with a particular break as well as price and volume graphs. The price and volume graph (see Figure 4) provides market

[7] CIA Server is an expert system suite from Haley Enterprises which implements RETE in a client-server architecture. Rules are authored in the Eclipse language, a descendent of CLIPS.

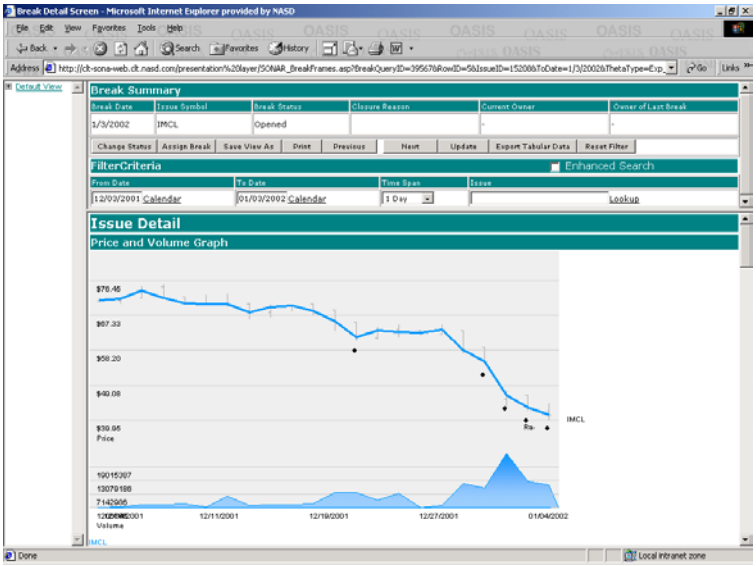

Figure 4: Issue Detail Display – Price and Volume Graph

data, indicators of breaks and text events, and news nature associated with any breaks in the issue. This gives user a quick overall view of the context surrounding a break.

Finally, SONAR has watch list feature which enables users to specify their own "text search agents" to provide routine search of incoming news and SEC filings. This is particularly valuable when specific individuals, companies, business sectors, or products are found to be "hot" areas for fraud.[8]

Database

SONAR's modules communicate via an Oracle database. The data model is designed for flexible representation of evidence and analyses – entities, relationships, and events are linked by easily generalized relations. It is anticipated that, as more post-detection modules are added to the system, this database will develop many of the aspects of a blackboard, including control and invoking of secondary analyses. This already occurs in the evaluation of broker volume concentrations when research reports are detected.

Some limitations to the generality of the model have been required due to performance considerations. These include a basic assumption of a "unit" of analysis – the daily features of a single security. We also assume an a priori universe of known security symbols and company names (roughly 100,000). Finally, since material news in insider trading cases is usually limited to PERM-R types, these are specifically modeled, rather than being represented in a more general E-R model.

Uses of Artificial Intelligence Technology

AI technologies and Domain Knowledge

The principal AI components of SONAR have been described above. NLP components include a rule-based component for mining information from text, fuzzy match for name consolidation, and topic tracking. Logistic regression models form the basis of both a learned detector of market events and a methodology for combination of data from disparate sources. Rule-based inference is being employed to refine and improve breaks after the initial detection.

The key to SONAR's success is its reliance on domain-specific knowledge. Text mining rules, word and concept lists, factors, theta equations, and inference rules are all carefully gleaned and maintained as part of a business process in place within the department. In fact, tuning of the statistical models by logistic regression is the only place where automated acquisition of knowledge takes place.

We considered several alternatives which would rely more heavily on learning technology, particularly in the NLP arena. However, the need to provide analysts, who must ultimately document their investigations, with understandable reasons for all system decisions led us to prefer a rule-based approach where one was feasible. In fact, Clear Forest's performance on information extraction benchmarks was comparable to the best competitors.

We had less clear a picture of the uses for the rule-based inference component, and chose CIA Server based on its implementation of RETE and the client-server architecture which promised performance and scalability. We have not yet been able to realize that performance.

Related Applications

SONAR derives a great deal from NASD's ADS system in terms of its role within the organizational work flow and its use of an institutionalized knowledge maintenance process. It is a Break Detection System as is ADS, providing NASD with the ability to do comprehensive surveillance of the entire market. However, SONAR more closely resembles some investigative and intelligence analysis systems in its dependence upon linkage, identification of entities, and intelligent fusion of data from different sources and of different forms. It incorporates consolidation and linkage features of the FinCEN AI System. [Senator 1995] The use of concepts extracted from text is related to capabilities in Document Explorer.

[8] A case in point was a search we undertook, early on, for post-911 bioterror products. In the aftermath of that tragedy and the subsequent anthrax attacks, a number of companies began touting "anti-bioterror" products in order to inflate the apparent value of their offerings. The search netted at least three companies which the SEC took to criminal court. The watch list feature was designed in part on our experiences with this search.

[Feldman 2002] Separate, trained detectors for sub-types of fraud are reminiscent of Activity Monitoring. [Fawcett 1999]

Application Use and Payoff

As a break detection system, SONAR is used to alert and augment human analysts in the task of detecting and evaluating potential cases of IT and Fraud. Thus, its performance should be evaluated as part of an overall process which includes alert detection, selection for review, and ultimately referral to prosecuting agencies.

By adding NLP analysis of news for materiality, using a more sophisticated approach of trending in determining profit potential or loss avoidance, and refining the use of factors by heuristics, SONAR has made significant improvements over its predecessor, SWAT. Table 1 compares both systems' results over 4 month periods, in terms of number of alerts generated and how many of them are then chosen for review and possible prosecution. A higher percentage of reviews shows reduction in false positive alerts and greater effectiveness of the system.

The total observed savings of roughly 6000 hours reviewing breaks, is borne out anecdotally. Annually, this equals a savings of 9 positions (out of about 30 in both groups). These resources were re-directed to greater detail in the later stages of review, resulting in more comprehensive and accurate regulation.

	SONAR	**SWAT (old system)**
Alerts (in 4 mo. period)	4,829	10,247
Reviews	180[9]	59
Percent	3.73%	0.58%
Time to review an alert	15-20 min	30-60 min
Volume concentration of brokerage firms	Provides	Doesn't provide
Material Events	Provides	Doesn't provide
SEC filings and misrepresented events.	Provides	Doesn't provide
History of alerts	Provides	Doesn't provide
Insider trading filings (Form -144)	Provides	Doesn't provide

Table 1: Comparison of SONAR and SWAT

SONAR also increases efficiency of review and investigation by providing references to the brokerage firms' volume concentration, SEC filings and other supporting information including key information from past cases. For most alerts, the only time an analyst needs to go outside SONAR is at the very last stage of the review process, called "bluesheeting", where the analyst collects customer account and other information on individuals directly from the company under review and brokerage firms who facilitated the trades. Table 1 compares the types of information provided in SONAR and SWAT.

Recalibrating Coefficients

To improve quality and maintain consistency with changing market conditions, selected breaks and near-breaks generated by SONAR are validated, on a weekly basis, by a core team of market analysts. Cases of wanted and of unwanted alerts are then used as training data to recalibrate the coefficients of each theta equation. Table 2 shows the outcome of one such recent calibration. 94% of the sample are classified correctly while both false positive and false negative rates are very low .[10]

	Non-Alert	Alert	Total
Unwanted	**105**	**7**	**112**
Wanted	**5**	**84**	**89**
Total	**110**	**91**	**201**

Table 2: Results of Theta Calibration

Application Development and Deployment

The SONAR project development team consisted of representatives from the Business Information Management, KDD, Insider Trading and Fraud teams of MRD, and software developers from the EDS Corporation, NASD's technology partner. At its peak the team consisted of approximately 35 people. Table 3 lists key SONAR development milestones.

September 1999	Study and analysis of material news announcements
October 1999	High level system design
August 2000	Rapid prototyping of text mining rule
September 2000	SONAR development
October 2000	Text mining results verification
January 2001	Study of market activity surrounding material news events
April 2001	Modeling of market activity and embedding news events in the model
December 2001	SONAR fully deployed
December 2002	New release with text mined evidence from EDGAR filings

Table 3: Application Development/Deployment

SONAR began as a proof-of-concept with the Insider Trading team. The project was initiated in September 1999 with a number of news story reading sessions to understand structure of the stories from various vendors

[9] The great increase in reviews is probably due in part to the period in question for SONAR corresponding to a significant downturn in the market, resulting in greater frequency of fraud and loss-avoidance IT.

[10] This is a significant improvement over SWAT, where the highest correct classification achieved was roughly 60%. This is due to the critical addition of material news.

and also to determine materiality, timeliness and uniqueness of stories. These sessions helped us defining the rules for timeliness and uniqueness, which were then used in rapid prototyping of the text mining rules.

Some of the modeling concepts from SWAT were used in developing models for SONAR. The results of these newly developed models were presented to the Insider Trading team via weekly domain meetings, beginning in January 2001, for verification.

Weekly domain meetings for Fraud section kicked off during April 2001 to understand the business knowledge of the Fraud section and their text mining and evidence gathering needs from EDGAR filings.

By the end of December 2001, SONAR, with its completed GUI interface, was deployed with some of the modeling work in progress and EDGAR evidence gathering work in study and analysis. In December 2002, a new release of SONAR was put in place to provide suspicious flags (evidence of misrepresentation) from EDGAR filings.

Knowledge Maintenance

Knowledge maintenance in SONAR is enabled by both tools and processes. Weekly meetings are held with key users in each domain area to review current breaks, text-mined evidence, post detection analysis results and status of new tasks. At these meetings, new scenarios are discussed and prototype models and rules are evaluated for inclusion in the system. Operational parameters are tuned to reflect the tradeoff between break quality and quantity consistent with the analysts' ability to evaluate them. As break quality improves, thresholds can be adjusted to allow more, marginal breaks, as well as allowing new types of violations to be detected. The KDD team also regularly analyze users' comments within evaluated breaks.

Future Directions

Future work on SONAR is aimed at two goals. (1) Provide greater flexibility to tailor the system to new domains, new markets and exchanges, and new regulatory concerns. This will be accomplished in part through even greater reliance on knowledge-based components. For example, new factors, and even new data sources should be able to be specified by description rather than hard-wired into software. Use of tools such as XML for component interfaces will help in this regard. (2) We want to enhance the post-detection analysis of breaks by automating more of the process of evidence gathering, linking, and re-analysis. Knowledge-based acquisition of secondary source data, refinement of models with knowledge extracted from prior breaks, and eventually an agent-like component to assist analysts in constructing an entire case are all envisioned.

Acknowledgments

We would like to thank all our colleagues at NASD and EDS who aided in the development of SONAR or contributed to its knowledge bases. Cam Funkhouser, Halley Dunn, Dave Katz, Jim Dolan, Maureen Siess, and their staff – our users and domain experts, were both patient and forward thinking. Darrin Hall, Jeff Schoppert, and Kathy Kidd of EDS were supportive and highly creative in providing guidance to the software development effort. We especially thank Steve Luparello, Executive VP Market Regulation, and Holly Lokken, VP Business Information Management, whose vision and support made this project possible.

References

Fawcett, Tom and Provost, Foster., "Activity Monitoring: Noticing Interesting Changes in Behavior," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pp. 53-62, ACM, 1999

Feldman, Ronen, "Document Explorer," in W. Kloesgen and J. Zytkow (eds.), *Handbook of Knowledge Discovery and Data Mining*, pp. 629-636, Oxford University Press, 2002.

Goldberg, Henry G. and Senator, Ted E., "Restructuring Databases for Knowledge Discovery by Consolidation and Link Formation," *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, Montreal, August, 1995.

Kirkland, James D., Senator, Ted E., Hayden, James J., Dybala, Tomasz, Goldberg, Henry G., and Shyr, Ping, "The NASD Regulation Advanced Detection System (ADS)," *AI Magazine* 20(1):55-67, 1999.

Senator, Ted E., Goldberg, Henry G., et. al., "The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions," *Proceedings of the Seventh Innovative Applications of Artificial Intelligence Conference*, Montreal, August, 1995.

Senator, Ted E., "Ongoing Management and Application of Discovered Knowledge in a Large Regulatory Organization: A Case Study of the Use and Impact of NASD Regulation's Advanced Detection System (ADS)," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-00)*, pp. 44-53, ACM, 2000.

Senator, Ted E. and Goldberg, Henry G., "Break Detection Systems." in W. Kloesgen and J. Zytkow (eds.), *Handbook of Knowledge Discovery and Data Mining*, pp. 863-873, Oxford University Press, 2002.