Indian Accounting Review
Vol. 20, No. 1, June 2016

Textual Analysis and Business Intelligence in Big Data Environment : Search Engine versus XBRL

Rajendra P. Srivastava

School of Business
University of Kansas, USA
E-mail: rsrivastava@ku.edu

ABSTRACT

The main objective of this paper is to discuss the role of search engines in data analytics within XBRL Environment and beyond. It is frequently argued that having XBRL formatted business reports makes it easy to access information and hence analyze it. However, one major problem with the XBRL formatted data is that the analytical tools will capture only those pieces of information that are tagged. What if the user needs information that is not tagged or not required to be tagged? Can the analytical tools still be effective to provide the analysis needed? This is where search engines would become important. The present paper highlights the importance of textual analysis and demonstrates the value of search engines that go beyond the XBRL environment, especially in situation where the needed pieces of information are not tagged. Several examples are presented in the paper to show the value of a search engine such as Seek iNF (*https://www.seekedgar.com*) developed at the University of Kansas, in performing textual analysis and developing predictive models, especially where no programming skills are required.

Key words: Search engine for SEC Filings, textual analysis, predictive models, Seek iNF.

I. INTRODUCTION

The main objective of this paper is to discuss the role of search engines in textual analysis, data analytics, and business intelligence. Traditionally, researchers have used financial and non-financial data to build models for analyzing financial risk and financial performance. For example, Altman (1968), Beaver (1996), Jones and Hensher (2004), Ohlson (1980), Shumway (2001), Zmijewski (1984) and many others have used financial data to develop models

Acknowledgements: The author would like to thank the participants of the 2014 International Symposium on Accounting Information Systems (organized by the Department of Accounting, University of Melbourne, Australia, January 2-3, 2014) for their valuable suggestions on an earlier version of the paper.

for assessing financial risk and financial performance using the regression technique. Emery and Cogger (1982) used cash flow data to develop a theoretical model to measure liquidity of a company, which in turn measures the financial risk or potential for bankruptcy. Several researchers have used the financial data to develop neural network models to forecast bankruptcy and fraud [e.g., see Boritz and Kennedy (1995), Boritz, Kennedy, and Albuquerque (1995), Fanning and Cogger (1998), O'Leary (1998)]. Lensberg, Eilifsen, and McKee (2006) have used financial data to develop bankruptcy prediction models using genetic algorithm, while McKee (2003) used similar data to develop bankruptcy prediction models using rough sets. Shumway (2001) used financial data to develop a simple hazard model for bankruptcy prediction.Thus, a vast amount of work has been done in using the financial data to develop predictive models and business intelligence for risk assessments and bankruptcy prediction.

Callen, Khan and Lu (2013) use financial data to measure "accounting quality" where "the precision with which financial reports convey information to equity investors about the firm's expected cash flows" defines the accounting quality. Recently, Lee, Churyk and Clinton (2013), (see also, Loughran and Mcdonald, 2011) developed a fraud detection model based on the counts of key words (positive versus negative words), and punctuation in Management Discussion and Analysis (MD&A) disclosures and found it to be performing better than many other models developed using financial data.

While financial and certain business data are being tagged using XBRL (Extensible Business Reporting Language), not all key words, punctuations, and present tense and past tense sentences, and counts of words and sentences that are needed for determining sentiments and readability indices to develop predictive models as developed by Lee, Churyk and Clinton (2013), are being tagged nor they would ever be tagged by XBRL technology. While there are many advantages of the XBRL technology from ease with which information can be retrieved, shared, transported across platforms, used in automated models, to cost saving and eliminating human errors in the above process (e.g., see, Srivastava 2009 and Debreceny et al. 2005), it is practically not possible to tag all words in a report through the XBRL technology. This is where the need of a text search engine becomes important, which is the topic of this paper. Text mining is a big area and an important area, especially in this era of big data. There are several books written on this topic. The focus of this paper is concentrated towards searching for financial and non-financial information, along with exploring the use of text analytics for predictive models in the context of accounting, auditing, and finance.

The Securities and Exchange Commission (SEC) in the USA and many similar government agencies in other countries (e.g., Securities and Exchange Board of India, SEBI) have mandated public companies to file their annual and quarterly reports in the XBRL format along with the text/html filings. Thus, all business relevant data, both financial and non-financial, are available in the XBRL tagged format. And therefore, it would be easy to create models that would provide timely information to analysts, regulators, and the general public

for business decisions as soon as the financial reports are filed with the regulatory agencies such as the SEC in the USA and SEBI in India. In fact, already several vendors have developed such analytical tools that use XBRL tagged data. Based on the availability of XBRL tagged business information from the SEC filings, the SEC has built a "RoboCop" (Novack, Carney, Harker 2013) to find irregularities in companies' filings. However, since the XBRL technology does not tag all the words, it is difficult to develop a model that would use the tagged financial and non-financial data along with the certain key words and number of or percentage of positive versus negative words in a document such as 10-K, the annual report filed with the SEC. In fact, SEC is looking into integrating text-based information into their "RoboCop" (Novack, Carney, Harker 2013), which is based on the tagged data.

Rest of the paper is divided into four sections. Next section provides a brief discussion on the background research, especially the textual analysis research. Section III describes the features of the search engine, Seek iNF,that has been developed at the University of Kansas. Section IV provides illustrations on how Seek iNF can be used to perform textual analysis without any programming skills in PERL or Python. Finally, Section V provides a summary and conclusion.

II. BACKGROUND ON TEXTUAL ANALYSIS RESEARCH

In this section, we briefly describe some of the important researches in the textual (i.e., content) analysis domain and show the importance of such work not only in developing predictive models for assessing financial risk, business risk, fraud risks, etc., but also in developing business intelligence to understand the quality of financial reporting and financial disclosures. Companies disclose certain piece of information, some voluntarily and some by mandate, in their management report to communicate to the public about what they have accomplished and what they plan to achieve in the future. The language how they communicate to public and regulatory agencies may vary from when the company is profitable versus when the company is having financial distress. Li (2010) provides a good review of the textual analysis related to corporate disclosures. Disclosures are becoming complex, as mentioned earlier, some due to regulatory requirements and some due to management intention as stated by Li (2010, p. 144):

> . . . managers' communication patterns could reveal certain managerial characteristics and thus have significant implications for understanding management decisions. Recent developments in behavioral economics emphasize the cognitive biases of human beings and the roles of these biases in decision making (Kahneman, 2003). . . As a communication vehicle for management, textual disclosures can provide a means for researchers to assess managers' behavioral biases and understand firm behavior.

In a recent article, Monga and Chasan (2015) find that :

> Companies are spending an increasing amount of time and energy beefing up their regulatory filings to meet disclosure requirements. The average 10-K is getting longer-about 42,000 words in 2013, up from roughly 30,000 words in 2000.

Tetlock, Saar-Tsechansky, and Macskassy (2008) examine whether the use of a simple quantitative measure of language can predict individual firms' accounting earnings and stock returns. They mention the following two advantages of using such a textual analysis (page 1438).

> First, by quantifying language, researchers can examine and judge the directional impact of a limitless variety of events, whereas most studies focus on one particular event type, such as earnings announcements, mergers, or analysts' recommendations. Analyzing a more complete set of events that affect firms' fundamental values allows researchers to identify common patterns in firm responses and market reactions to events. Equally important, examining all newsworthy events simultaneously limits the scope for "dredging for anomalies"- the phrase used by Fama (1998) to describe running event studies on different types of events until one obtains "significant" results.
>
> Second, linguistic communication is a potentially important source of information about firms' fundamental values. Because very few stock market investors directly observe firms' production activities, they get most of their information secondhand. Their three main sources are analysts' forecasts, quantifiable publicly disclosed accounting variables, and linguistic descriptions of firms' current and future profit-generating activities. If analyst and accounting variables are incomplete or biased measures of firms' fundamentals, linguistic variables may have incremental explanatory power for firms' future earnings and returns.

Loughran and Mcdonald (2011) analyze a large sample of 10-Ks during 1994-2008, and show that almost three-fourths of the word count identified as negative by the Harvard Dictionary do not have negative meaning in a financial context. In addition, they developed an alternative negative word list that better reflects the tone of financial text along with five other word classifications (positive, uncertain, litigious, strong modal and weak modal). They linked the word lists to 10-K filing returns, trading volume, subsequent return volatility, fraud, material weakness, and unexpected earnings.

Lee, Churyk and Clinton (2013, p. 35) state that "Conventional fraud detection measures using ratio analysis and other financial data were either unable to detect the fraud or unable to detect it soon enough to avoid catastrophic outcomes". They develop a fraud detection model based on the counts of key words (positive versus negative words), and punctuation in Management Discussion and Analysis (MD&A) disclosures and found it to be performing better than many other models developed using financial data.

In a preliminary investigation, they found that companies that were involved in fraudulent financial reporting used on average fewer negative words compared to similar companies. Their early fraud detection model based on the textual (i.e., content) analysis of MD&A in 10-K is:

FRAUDi = 2.89757 – 0.83408 (POSITIVE EMOTIONi) – 0.48315 (PRESENT TENSEi) + .0001 (TOTAL WORDSi) – 2.80753(COLONSi)

While Lee, Churyk and Clinton (2013) validate their model empirically, the model makes logical sense and conveys about managements' behavior when they commit fraud. It seems the Management is trying to misguide the users of the financial report and hide the fraud by using more words, fewer present tense, less 'positive emotion' and fewer colons in the management discussion and analysis part of 10-K.

Li, Lundholm, and Minnis (2013) develop a model to compute management's perception of the intensity of competition using textual analysis of firms' 10-K filings. They find that the measure of competition varies both across-industry and within-industry, and also show that each measure is related to the firm's future rates of diminishing marginal returns. Their measure is based on the count of the number of occurrences of words "competition, competitor, competitive, compete, competing," including those words with an "s" appended, less any case where "not," "less," "few," or "limited" precedes the word by three or fewer words. Their measure of competition is:

PCTCOMP = 1000*NCOMP/NWORDS,

where NCOMP = number of words in 10-K as described above and NWORDS = Total number of words without numbers. They wrote their own program to count the occurrences of these words and phrases because there was no database that would provide such counts. However, recently, the search engine, Seek iNF, developed at the University of Kansas, has several built-in features that provides such counts without any required programming skills (for details see the web site *https://www.SeekEdgar.com:8443/*). This will be further elaborated in the next section.

One can develop predictive models and business intelligence using the textual analysis of a report along with the financial and other non-financial data for assessing risks such as business risk, litigation risk, and the risk of financial distress. But to do that one needs effective text mining tools and search engines. There are several such tools available in the market place from providers such as LexisNexis (*http://www.lexisnexis.com/hottopics/lnacademic/*) and IBM (*http://www-03.ibm.com/software/products/en/category/content-analytics*). However, rather than describing all different search engines and text analyzing systems, I describe, next, the search engines, Seek iNF, that provides tremendous opportunity not only to perform textual analyses and develop predictive models for assessing various risks, such as, financial risk, litigation risk, and fraud risk, but also to create new knowledge.

III. SEEK iNF – A TOOL FOR INFORMATION RETRIEVAL AND TEXTUAL ANALYSIS

As discussed in the previous section, textual analysis of annual reports and other companies' filings is playing an important role in developing predictive models and business intelligence. However, the lack of Perl and Python programming skills in business faculty has restricted the research in textual analysis to very few, as evidenced by the published research. For example, Li, Lundholm, and Minnis (2013) wrote their own program to extract the information they needed to conduct their research.

In this section, I plan to describe a search engine, Seek iNF (Search Engine for Extracting Knowledge from Industrial Filings) that does not require any programming skills to extract information from any report or filing whether it is in the traditional text format, or any other format such as HTML, ASCII, PDF or XBRL. Seek iNF is an outgrowth of a project FRAANK (Financial Reporting and Auditing Agent) that started in 1996 at the Ernst & Young Center for Auditing Research and Advanced Technology in the School of Business at the University of Kansas (see Bovee et al. 2005 for details about FRAANK). The basic idea was to develop an intelligent system that will fetch a report from a website and parse the text to extract the desired information. FRAANK succeeded, in no time, in fetching 10-K (annual reports) and 10-Q (quarterly reports) of the US public companies from the SEC Edgar database and parsing the financial information from the balance sheet, income statement, and cash flow statement and dumping it into MS Excel Spreadsheet. The University of Kansas received the US Patent for FRAANK in 2010.

FRAANK did not have any intelligence in terms of getting specific piece of information from the reports. Once FRAANK became functional, my colleagues started requesting for specific piece of information such as 'a list of all the companies that have mentioned this issue ... in the footnote in 10-K'. We will tweak the program in FRAANK and run it on the weekend and have the data ready by Monday. Next, an idea was suggested that why not we develop a system where users query the system and get the data they want on their own without requesting us. This is when Seek iNF was born in 2007.

Seek iNF is a search engine based on the Cloud technology. There is no need to have any programming skills to extract information from a report that is in the database. Currently, Seek iNF uses its own database of SEC Filings and Public Company Accounting Oversight Board (PCAOB) inspection reports that are pre-processed for efficient retrieval of information. At the present, this database consists of 14 million filings and 20 million documents for the years 1994 – 2016 and daily updated automatically (*see* Table 1 for details). It has built-in features for identifying exact phrases containing words like articles and prepositions. In addition, it has built-in Boolean logic, counters of words, phrases, and sentences, and provides distribution of all the words in a document and six readability indices, all without any need or programming.

TABLE 1

List of Filings and Exhibits in the Seek iNF Database

File Type *Or* Exhibit Type	**Mean**
AAERS	Accounting and Auditing Enforcement Releases
6-K, 6-K/A	Current report of foreign issuer pursuant to Rules 13a-16 and 15d-16 Amendments
8-K, 8-K/A	Current report filing (Disclosure of significant economic events), and amendment
10-K, 10-K/A	Annual report pursuant to Section 13 and 15(d), and amendment
10-KT, 10-KT/A	Annual-Transition report pursuant to Rule 13a-10 or 15d-10, and amendment
10-K405, 10-K405/A	Annual Reports – Reg. S-K Item 405, and amendment
10-KT405, 10-KT405/A	Annual-Transition Reports - Reg. S-K Item 405, and amendment
10-KSB, 10-KSB/A	Annual Reports by Small Businesses, and amendment
10-KSB40, 10-KSB40/A	Annual Reports by Small Businesses – Reg. S-B Item 405, and amendment
10-Q, 10-Q/A	Quarterly report pursuant to Section 13 or 15(d) and amendment
10-QT, 10-QT/A	Quarterly-Transition report pursuant to Rule 13a-10 or 15d-10 and amendment
10-QSB, 10-QSB/A	Quarterly Reports by Small Businesses and amendment
13D,13D/A (Schedule)	Schedule filed to report acquisition of beneficial ownership of 5% or more of a class of equity securities
13G, 13G/A (Schedule)	Schedule filed to report acquisition of beneficial ownership of 5% or more of a class of equity securities by passive investors and certain institutions
13F-HR, 13F-HR/A	Initial Quarterly Form 13F Holdings Report filed by institutional managers. Initial Quarterly Form 13F Combination Report filed by institutional managers
13F-NT, 13F-NT/A	Initial Quarterly Form 13F Notice Report filed by institutional managers
20-F, 20-F/A	Annual and transition report of foreign private issuers pursuant to Section 13 or 15(d)
40-F, 40-F/A	Annual reports filed by certain Canadian issuers pursuant to Section 15(d) and Rule 15d-4
424B1 - 424B8	Prospectus filed pursuant to Rule 424(b)(1) – Rule 424(b)(8)

TABLE 1 (*Contd.*)

File Type *Or* Exhibit Type	**Mean**
N-CSR, N-CSR/A	Certified annual shareholder report of registered management investment companies filed on Form NCSR
N-CSRS, N-CSRS/A	Certified semi-annual shareholder report of registered management investment companies filed on Form NCSR
CORRESPRESPONSE	A correspondence can be sent as a document with another submission type or can be sent as a separate submission. Response by SEC to the above correspondence
DEF 14A, DEFA 14A	Definitive proxy statements and amended proxy statement
Exhibit-21 (10-K)	List of Subsidiaries
Exhibit-95 (10-K)	Mine Safety Disclosure
Form 3, Form 3/A (F-3, F-3/A)	Registration statement for specified transactions by certain foreign private issuers
Form 4, Form 4/A (F-4, F-4/A)	Registration statement for securities issued by foreign private issuers in certain business combination transactions
Form 5, Form 5/A (F-5, F-5/A)	Annual statement of changes in beneficial ownership of securities
S-1, S-1/A	General form of registration statement for all companies including face-amount certificate companies
PCAOB Inspection Reports	PCAOB reports inspecting registered public account ing firms to assess compliance with the Sarbanes-Oxley Act, the rules of the Board, the rules of the Securities and Exchange Commission, and professional standards, in connection with the firm's performance of audits, issuance of audit reports, and related matters involving U.S. companies, other issuers, brokers and dealers.
PCAOB Settled Disciplinary Orders	Board Orders in Settlements with Registered Firms or their Associated Persons
PCAOB Adjudicated Disciplinary Actions	Opinions, Orders, and Other Final Board Action Imposing Sanctions in Contested Disciplinary Proceedings and SEC Actions on Review of those Sanctions

Basically, Seek iNF deals with the following four Dimensions: (1) Search all or few documents in the database for specific issues or concerns, (2) Obtain a piece of information whether financial or non-financial, (3) Perform text analytics, and (4) Download all the searched data in HTML document and Excel Spreadsheet file for further analysis. In the next section, I will describe these

dimensions in detail with examples. These four dimensions along with the following Power Features, Seek iNF provides unique opportunity to researchers to perform textual analysis and develop predictive models and business intelligence in an efficient and effective manner which has not been possible because of lack of programming skills in researchers except a few.

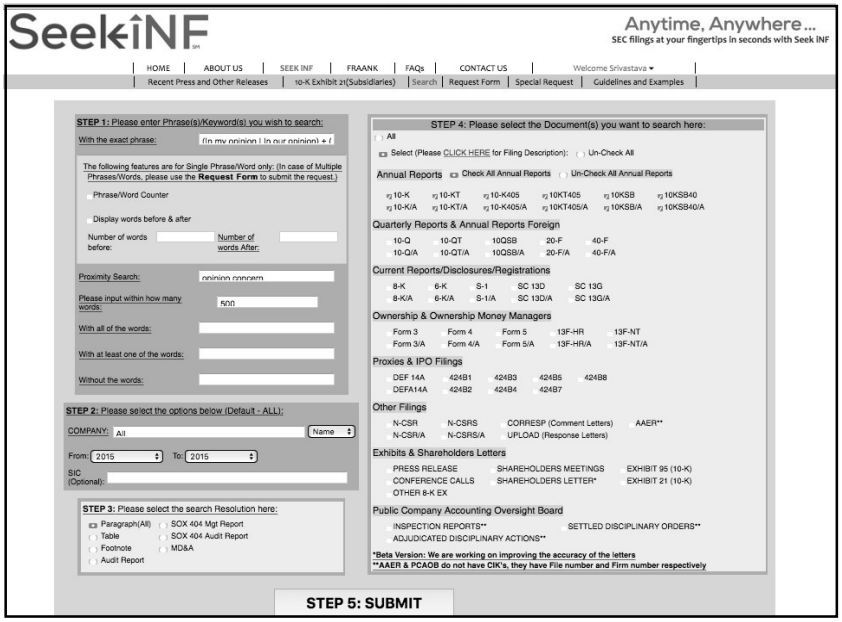
Seek iNF has several Power Features that make the search and retrieval of information much more efficient and effective which has never been possible without the knowledge of programing by researchers. These features are discussed next.

Search Documents with Multiple Exact Phrases: This feature allows users to search all or few documents in the database for presence or absence of multiple phrases containing words like 'a', 'in', 'if', 'no', 'of', 'the', etc. Such a feature makes it easier to find which of the companies have mentioned in their 10-K a given phrase, say 'wrongful termination'. One can input this phrase in the search engine and get 4,511 companies in few seconds that have mentioned the phrase in their 10-Ks. I will elaborate on this feature further through examples in the next section. Using the built-in Boolean logic makes the search process much more effective, especially using the multiple exact phrases.

You can search any document with a combination of exact phrases with "AND", "OR", or any combination of these conditions. Seek iNF uses "+" for AND, "|" for OR, and "-" for negation with or without space before and after the symbol. Here are examples of how one would type the three phrases represented by A, B, & C in the exact phrase slot in Step 1 (*see* Figure 1).

FIGURE 1

Screenshot of the main page of Seek iNF



- For AND logic, type the phrases separated by + sign as: A + B + C
- For OR logic, type the phrases separated by | symbol as: A | B | C
- For AND and OR, such as A & (B or C), type A + (B | C)
- For AND andnegation such as "A & B & not C", type as A + B – C
- For OR and Negation such as "(A or B) & not C", type as (A | B) – C

Information Retrieval with Proximity Search: This feature searches documents for information with two words within any number of words. The number of words between two words could be anywhere between two to several thousands. This is very useful in searching for information; especially when the two words of interest are separated by few words in the text. For example, suppose you are interested in finding executive bios. This information is usually published in the proxy statement (DEF 14A) usually in a table. In order to get this table, one can use the Proximity search feature with two words 'name' and 'age' within two words. You will get all the tables in a few seconds. This example is further elaborated in the next section.

Information Retrieval with few words before and few words after a phrase: This feature provides users a tool to obtain a unique piece of information along with the traditional data that are available through a database. For example, one can find addresses of companies or fiscal year end using this feature, which is not currently available through any database provider. Several examples are used to illustrate this feature in the next section.

Counts of Multiple words, phrases, and sentences: This feature provides a tool to gather counts of multiple words, multiple phrases, and sentences, along with distribution of all the words. This kind of information is useful in textual analysis and developing predictive models and business intelligence. This will be further elaborated in the next section through examples.

Counts Occurrence of two words within few words: This feature is useful again in developing models for business intelligence. For example, one can use this feature to compute the metric measuring competition as developed by Li et al. (2013). I will further elaborate this feature through examples in the next section.

Readability Indices: Seek iNF has been programmed to provide the following six readability indices: Gunning-Fog Index, Smog Index, Flesch Reading Ease, Flesch-Kincaid Grade Level, Automatic Readability Index, Coleman-Liau Index (*see* Table 2 for definition), for all the documents present in the Seek iNF database. This means these measures are available for all the SEC Filings and PCAOB Inspection Reports (about 20 million documents). With this feature one can perform textual analysis on all the filings and documents without knowing any programming.

TABLE 2

Readability Indices

Readability Index	**Formula**
Gunning-Fog Index	0.4x[(words/sentences) + 100x(complex words)/ words][a]
Smog Index	1.0430xSQRT[(number of polysyllables)x30/ (number of sentences)] + 3.729[b]
Flesch Reading Ease	206.835 – 1.015[(total words)/(total sentences)] – 84.6[(total syllables)/(Total words)][c]
Flesch-Kincaid Grade Level	0.39[(total words)/(total sentences)] + 11.8[(total syllables)/(total words)] – 15.59[c]
Automatic Readability Index	4.71(characters/words) + 0.5(words/sentences) – 21.43[d]
Coleman - Liau Index	0.0588L – 0.2965S – 15.8[e]

a. *https://en.wikipedia.org/wiki/Gunning_fog_index*. A "complex word" is defined as the word with three or more syllables. Do not include proper nouns, familiar jargon, or compound words. Do not include common suffixes (such as -es, -ed, or -ing) as a syllable.
b. *https://en.wikipedia.org/wikiSMOG*.
c. *https://en.wikipedia.org/wiki/Flesch%E2%80%93 Kincaid_read-ability _tests*.
d. *https://en.wikipedia.org/wiki/Automated_readability_index*
e. *https://en.wikipedia.org/wiki/Coleman%E2%80%93Liau_index.* L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

Download all the searched data in Excel Spreadsheet: This feature allows users to download all the output information in a CSV file and also in an HTML file, irrespective of whether the searched output is in the form of HTML snippets, or tables, or whether it is the counts of multiple phrases/ words. A CSV file can be opened in Excel and thus makes it easier to perform statistical analysis of the output data.

In addition to the above features, Seek iNF provides the frequency distribution of all the words in a document which can be downloaded in an Excel file for textual analysis. The users can search for specific piece of information by companies' names or CIKs (Central Index Keys by SEC), or by SIC (Standard Industrial Classification). Also, one can search for information by specific resolution such as "paragraph", "Table", "Footnote" "Audit Report", "SOX404 Management Report", "SOX 404 Audit Report" and "MD&A".

IV. ILLUSTRATIONS

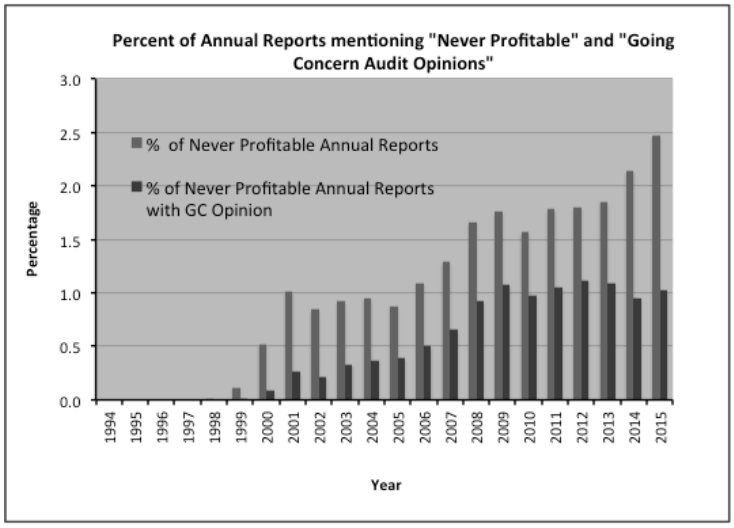
In this section, I demonstrate the power of the search engine, Seek iNF, in research, especially in accounting, finance, and other business disciplines.Given the ease with which one can gather information using Seek iNF, the knowledge, imagination, and creativity of the researcher will play an important role in future research.

Example 1: Searching for documents containing the exact phrases

Did you know that there are public companies that make statements in their annual reports (10-Ks) that they were not profitable and may never be profitable and still investors invest money? Why is that? How would you find out who they are? Manually, it is impossible! But, Seek iNF can answer this question in seconds. Insert the following phrases "We may never become profitable | We May Never Achieve or Sustain Profitability | we may never achieve or maintain profitability | We may never be profitable" in the exact phrase slot, and select "all companies", the time period, and 10-Ks, and get the result in seconds (*see* Figure 1). Figure 2 shows that such companies started around 1998 and has been growing in number since then. One can explore several questions as to not only who these companies are but to how can they survive so long even after making such a statement. Some have survived 10-12 years.

FIGURE 2

Graph of Percentage of Annual Reports with ‘‘Never Profitable’’and Going Concern Opinion by Year



Another interesting question comes to mind is what type of audit opinions were issued by the corresponding auditors. You would expect all such companies should have received a Going Concern audit opinion. This is easy

to find out using Seek iNF. Just insert the phrases pertaining to 'we may never be profitable' and pertaining to 'Going Concern opinion' together in the slot of "With exact phrase" in Step 1. More specifically, type the following phrases "(we may never become profitable | we may never achieve or sustain profitability | we may never achieve or maintain profitability | we may never be profitable) + (In my opinion | In our opinion) + (substantial doubt about| substantial doubts about | substantial doubts regarding | substantial doubt regarding) + going concern[1]" in the "With exact phrase" slot and the two words: "opinion concern", in the "Proximity Search" slot within 500 words and select "all companies", the time period, and 10-Ks. Figure 2 shows an interesting result - not all such companies received GC opinions. Actually, less than 50% have received a going concern opinion. Why is that? What about audit quality? Many interesting research questions emanate from this result.

Example 2: Proximity Search

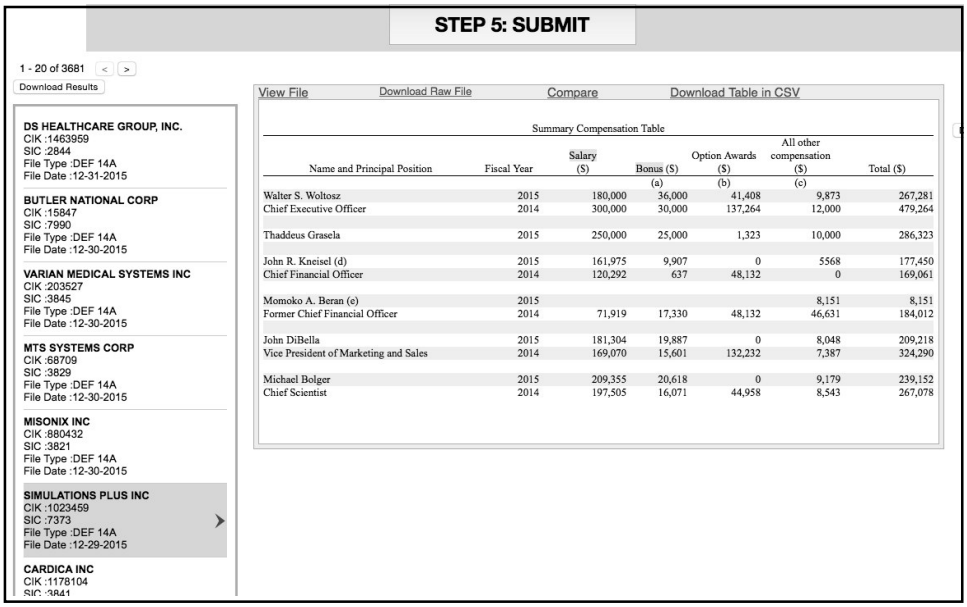
This feature also provides a powerful tool to capture useful information that is not available from any other source. Here I will demonstrate two examples that I have been asked frequently. One deals with executives' compensation and the other deals with executives' bios. My purpose here is to demonstrate the process. As a researcher one needs to identify what combinations of words and phrases will give exactly what one wants for one's research using this feature or in combination of other features.

2(a). Find Executive Compensation Tables. The executive compensation data are presented in a table in the proxy statement (DEF 14A). After looking at few compensation tables, I found out that the two words, 'salary' and 'bonus', were appearing within two words in almost all the tables in DEF 14As. Suppose, I want to obtain the executive compensation data for all the companies for the year 2015. Here is what I needed to do. Type in "salary bonus" in 'Proximity search', within 2 words, select "All" companies, the year 2015, 'Table' as the resolution, and DEF 14A as the filings, and submit the query. In few seconds, the system will return the result, with a displayof first 20 companies on the left of the display window. In the middle, the system will show the compensation table of the first company. Figure 3 is a screenshot of the compensation data displayed by the systemin the middle for one of the companies. One can download the compensation table displayed in the middle in the Excel Spreadsheet by selecting "Download Table in CSV" from the menubar on the top right side of the display window. One can download the entire searched result, i.e., all the tables in the CSV format by submitting a Request Form. The system will automatically send you the link for the data when it is ready (*see* the details at: *https://www.seekedgar.com:8443/SeekiNF_search_Engine.pdf*).

[1] Seek iNF uses '|' symbol for OR logic, '+' symbol for AND logic, and '–' for negation.

FIGURE 3

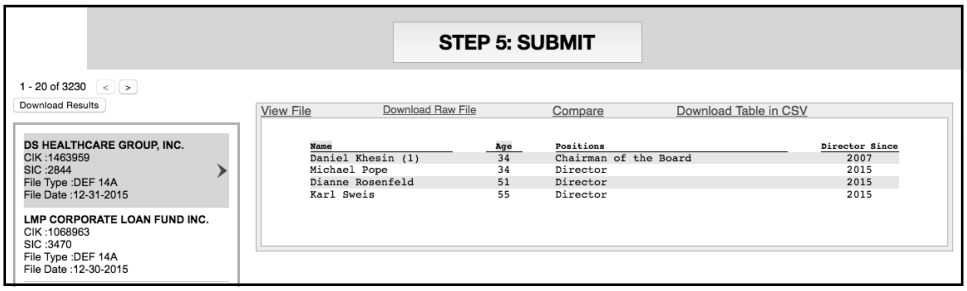
Screenshot of Compensation Table



2(b). Find Executive Bios from DEF 14A. Similar to the previous example of executive compensation, executive bios are in a table in the proxy statement (DEF 14A). Again one needs to find the pattern, which will identify such a table. After looking at few tables that contained names and ages of executives, I found out that 'Name' and 'Age' were appearing within two words almost in all such tables. Next, I used the two words 'Name' and 'Age' in "Proximity search" within 2 words to get all the tables that contained executives' bios for all the companies. Figure 4 is a screenshot of the display window for one of the companies. One can download this table in Excel by selecting "Download Table in CSV" from the menu bar displayed on the top right side of the display window. One can download the entire search result in Excel by submitting a Request Form and filling out the search criteria again. The system will automatically process the request and inform the user with the link when the data are ready to be downloaded.

FIGURE 4

Screenshot of Seek iNF Output of Executive Bios



Example 3: Search with few words before and few words after a phrase

This feature of searching for information with few words before and few words after a phrase is again a pretty powerful tool for getting information that is not available by any other source unless you know how to program in PERL or Python to fetch that information. Suppose you want to find the "fiscal year end" of all the public companies filing with the SEC. This information is provided by the companies in their 10-Ks, the annual reports filed with the SEC. Looking at few 10-Ks, it seems that this information is listed right after the phrase "fiscal year end". Thus, if one uses the search criteria "fiscal year end" in"With exact phrase" and selects the option to display the searched data with zero word before and 1 word after, the system will display the desired result. Figure 5 represents the "fiscal year end" data in an Excel file for a set of companies obtained through the process of submitting a "Request Form" (*https://www.seekedgar.com:8443/SeekiNF_search_Engine.pdf*).

FIGURE 5

''Fiscal year end'' data in Excel for a given set of Companies using the feature ''Proximity search'' and ''Zero Word Before and One Word After''

	A	B	C	D	E	G
1	CIK	COMPANY NAME	SIC	FILING DATE	FILE TYPE	
2						
3	1057060	MARINEMAX INC	5531	12/8/15	10-K	FISCAL YEAR END: 0930
4	1373690	AMERICAN PARAMOUNT GOLD CORP.	1000	12/8/15	10-K	FISCAL YEAR END: 0831
5	1429764	Car Charging Group, Inc.	3612	12/8/15	10-K	FISCAL YEAR END: 1231
6	788329	JOHNSON OUTDOORS INC	3949	12/8/15	10-K	FISCAL YEAR END: 0930

Example 4: A Measure of Competition Based on Textual Analysis of 10K

As mentioned earlier, Li, Lundholm, and Minnis (2013) compute management's perception of the intensity of the competition using textual analysis of the firm's 10-K filing. They count the number of occurrences of the following words: competition, competitor, competitive, compete, competing, including those words with an "s" appended, and then remove any case where "not," "less," "few," or "limited" precedes the word by three or fewer words.

They used their own programming skills to get these counts. However, I want to show here how one can get the required counts easily using the Seek iNF search capabilities without having any knowledge of PER or Python. As given earlier, their measure of competition is

PCTCOMP = 1000*NCOMP/NWORDS,

where NCOMP = number of words in 10K as described above and NWORDS = Total number of words without numbers. Seek iNF yields these counts in no time. We obtain NCOMP in two steps. First, we count the occurrence of the words: competition, competitor, competitive, compete, competing, competitions, competitors, competes, and subtract from it the "Proximity" count which counts the occurrence of the following two words within three or less words: not competition, less competition, few competition, limited competition, not competitor, less competitor, few competitor, limited competitor, not competitive, less competitive, few competitive, limited competitive, not compete, less compete, few compete, limited compete, not competing, less competing, few competing, limited competing, not competitions, less competitions, few competitions, limited competitions, not competitors, less competitors, few competitors, limited competitors, not competes, less competes, few competes, limited competes.

Let us compute PCTCOMP for the following five companies: Qwest Corp, Verizon Communications Inc, AT&T Inc., Level 3 Communications Inc., General Communication Inc., with CIKs: 68622, 732712, 732717, 808461, 794323, for 10 years (2006-2015). Since "Request Form" will allow you to download only five years data at a time, you need to submit two separate requests. After typing your name, email and University/Company, select the menu item "Phrase(s)/ Word(s) Count" for the words counts, and "Proximity Count" for the second part of the count to be subtracted from the first count to determine NCOMP. Seek iNF provides the total word count as a default. Thus, we can easily calculate the competition metric PCTCOMP.

FIGURE 6

Changes in the Competition Metric for Five Companies (ATT&T Inc., General Communication Inc., Level 3 Communications Inc., Qwest Corp, and Version Communication Inc.) for the Years 2005–2015

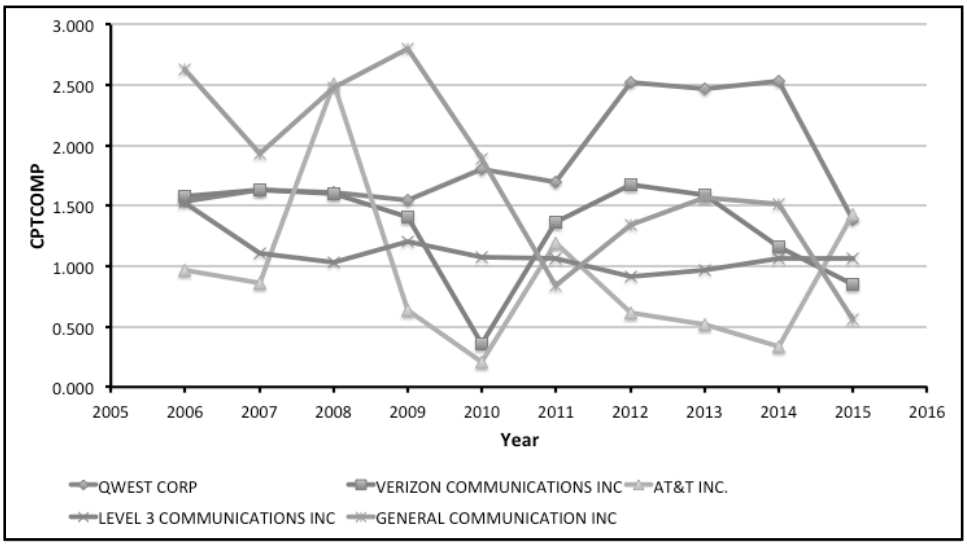


Figure 6 depicts a graph of the competition metric, PCTCOMP, for five companies in the telecom industry. As you can see Level 3 has had very little variation in its measure of competition where as for Verizon and AT&T it varies a lot during this period. There are a lot of research questions that one could explore about the economic reasons for these variations in the competition metric for individual companies. I do not explore these questions and try to find answers. I recommend this to be future research questions. One can perform all kinds of textual analyses and readability analyses on all the 14 million SE Filings and 20 million documents in a matter of minutes.

Example 5: Unique Financial Information from 10K

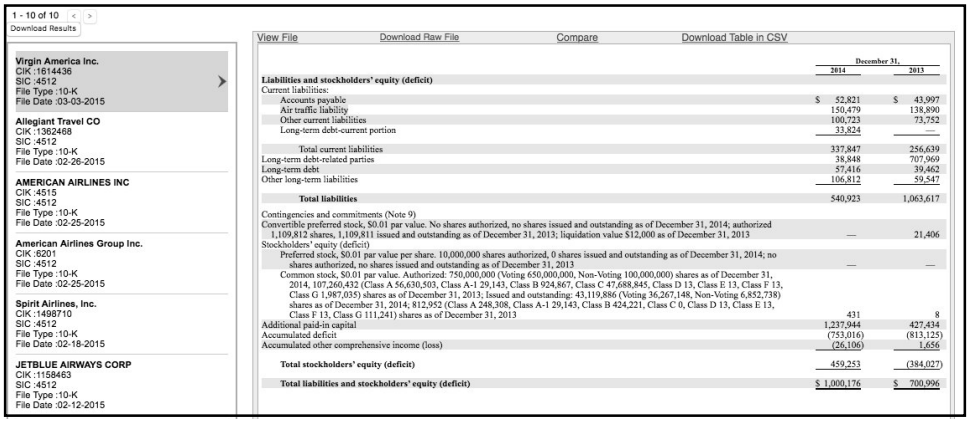
There are many unique financial line items that are not available in any of the current databases. For example, suppose you want to find "Air Traffic Liability" for airline companies for the year 2015. Here is what one can do to get this information through Seek iNF:Step1: Type the phrase Air Traffic Liability in the slot "With the exact phrase". Step 2: Since we want all the companies, leave the default choice "All" and select the desired time frame 2015-2015.Step 3: Select "Table". Step 4: Select 10-K. Step 5: Click on SUBMIT button.

You can download all the data in Excel Spreadsheet two ways, one directly from the display window by selecting the menu item "Display Table in CSV" just above the display window on the right for each company, and the other

by submitting a "Request Form" as done in Example 4. Figure 7 is a screenshot of the display window for one company. One can download this table in Excel too as mentioned above.

FIGURE 7

Unique Line Item, ‘‘Air traffic Liability’’ in Balance Sheet



V. SUMMARY AND CONCLUSION

This paper discusses the role of search engines in textual analysis and in developing predictive models and business intelligence in a big data environment and within XBRL environment. The paper first provides a summary of the background research on textual analysis and predictive models already developed and their importance in the current environment. It seems if managements want to misguide the users of the financial reports and hide the fraud, they would use more words, fewer present tense, less 'positive emotion' and fewer colons in the management discussion and analysis part of the 10-K as observed by Lee, Churyk and Clinton (2013). Although XBRL technology is being required by regulators globally for financial reporting purposes because of certain advantages, this technology lacks in providing the kind of data needed for textual analyses. Thus, search engines specifically designed to search textual information become important. Seek iNF is such a search engine that has been developed at The University of Kansas and is being further developed by SeekEdgar LLC (see *https://www.seekedgar.com* for details). Several examples have been discussed to illustrate the power of Seek iNF.

The text parsing search engines with built-in intelligence for searching for specific piece of information, financial and non-financial, is going to play an important role in future research for developing predictive models and business intelligence. I would emphasize that the knowledge, imagination and creativity of the researcher would drive the future research; not the pre-canned data.

References

Altman, E. (1968). Financial ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy.Journal of Finance 23: 589-609.

Atiya, A. F. (2001). Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results. IEEE Transactions on Neural Networks, Vol. 12, No. 4, July: 929-935.

Beaver, R. (1996). Financial ratios as predictors of failure.In Empirical Research in Accounting: Selected Studies 1966, Journal of Accounting Research, vol. 4, pp. 71-111.

Boritz, J. and D. Kennedy. (1995). Effectiveness of neural network types for prediction of business failure.Expert SystemsApplications, vol. 9: 504-512.

Boritz, J., D. Kennedy, and A. Albuquerque. (1995). Predicting corporate failure using a neural network approach.Intelligent System in Accounting, Finance, and Management, Vol. 4: 95-111.

Bovee, M., A. Kogan, R. P. Srivastava, M. A. Vasarhelyi, K. M. Nelson, (2005). Financial Reporting and Auditing Agent with Net Knowledge (FRAANK) and eXtensible Business Reporting Language (XBRL). Journal of Information Systems, Vol. 19, No. 1 (Spring): pp. 19-41.

Callen, J. L. M. Khan and H. Lu. (2013). Accounting Quality, Stock Price Delay, and Future Stock Returns. Contemporary Accounting Research, Vol. 30, Issue 1, Spring: 269-295.

Debreceny, R. S., A. Chandra, J. J. Cheh, D. Guithues-Amrhein, N. J. Hannon, P. D. Hutchison, D. Janvrin, R. A. Jones, B. Lamberton, A. Lymer, M. Mascha, R. Nehmer, S. Roohani, R. P. Srivastava, S. Trabelsi, T. Tribunella, G. Trites, and M. A. Vasarhelyi. (2005). Financial Reporting in XBRL on the SEC's EDGAR System: A Critique and Evaluation. Journal of Information Systems 19 (2):191-210.

Emery, G. and K. Cogger. (1982). The Measurement of Liquidity. Journal of Accounting Research, Vol. 20, No. 2 Pt. I Autumn: 290-303.

Fanning, K. M. and K. O. Cogger. (1998). Neural network detection of management fraud using published financial data. International Journal of Intelligent Systems in Accounting, Finance & Management, Vol. 7, No. 1, March: 21-41.

Jones and Hensher. (2004). Predicting Firm Financial Distress: A Mixed Logit Model. The Accounting Review, Vol. 79, No. 4, 2004, pp. 1011-1038.

Lee, C., N. T. Churyk, and B. D. Clinton. (2013). Detect Fraud Before Catastrophe. Strategic Finance, March: 33-37.

Lensberg, T., A. Eilifsen, and T. E. McKee. (2006). Bankruptcy theory development and classification via genetic programming. European Journal of Operational Research, Vol. 169, Issue 2, March: 677-697.

Li, F. (2010). Textual Analysis of Corporate Disclosures: A Survey of the Literature. Journal of Accounting Literature, Vol. 29: 143-165.

Li, F., R. Lundholm, and M. Minnis. (2013). A Measure of Competition Based on 10-K Filings. Journal of Accounting Research, Vol. 51, No. 2 (May): 399-436.

Loughran, T. and B. Mcdonald. (2011). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, Vol. 6, Issue 1, February: 35-65.

McKee, T. E. (2003). Rough sets bankruptcy prediction models versus auditor signaling rates. Journal of Forecasting, Vol. 22, Issue 8, December: 569-586.

Monga, V. and E. Chasan. (2015). The 109,894-Word Annual Report: As regulators require more disclosures, 10-Ks reach epic lengths; how much is too much? The Wall Street Journal, Business CFO Journal. Updated June 1.

Novack, J., J. Carney, and F. Harker. (2013). How SEC's New RoboCop Profiles Companies For Accounting Fraud.http://www.forbes.com/sites/janetnovack/2013/08/09/how-secs-new-robocop-profiles-companies-for-accounting-fraud/

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research 18: 109-131.

O'Leary, D. (1998). Using Neural Networks to Predict Corporate Failure. International Journal of Intelligent Systems in Accounting, Finance & Management, Vol.7, pp. 187-197.

Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. The Journal of Business, Vol. 74, No. 1, January: 101-124.

Srivastava, R. P. (2005). Financial Reporting in XBRL on the SEC's EDGAR System: A Critique and Evaluation. Working Party of the AAA Information Systems and Artificial Intelligence/Emerging Technologies Section (with 17 other members, with equal participation). Journal of Information Systems, Vol. 19, No. 2, Fall: 191-210.

Srivastava, R. P., (2009). XBRL (Extensible Business Reporting Language): A Research Perspective. Indian Accounting Review, Vol. 13, No. 1, pp. 14-32.

Tetlock, Paul C., M. Saar-Tsechansky, and S. Macskassy, (2008). More than words: Quantifyinglanguage to measure firms' fundamentals, Journal of Finance 63, 1437-1467.

Zmijewski, M. (1984). Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research 22 (Supplement): 59-82.