The Spanish Review of Financial Economics 10 (2012) 1–10

The Spanish Review of Financial Economics

www.elsevier.es/srfe

Article

Crawling EDGAR

Diego García [a,*], Øyvind Norli [b]

[a] University of North Carolina at Chapel Hill, United States
[b] BI Norwegian Business School, Norway

ARTICLE INFO

Article history:
Received 28 March 2012
Accepted 5 April 2012
Available online 9 May 2012

JEL classification:
G14
G18
G32

Keywords:
EDGAR
8-K statements
Text analytics
Executive turnover

ABSTRACT

While the title may lead you to think that this paper is about spiders, it is about firms in the United States reporting relevant business information to the Securities and Exchange Commission (SEC). The paper is meant to serve as a primer for economists in the computing details of searching for information on the Internet. One important goal of the paper is to show how simple open-source computer scripts can be generated to access financial data on firms that interact with regulators in the United States.

1. Introduction

Business relevant information is more easily available today than ever before. Information about corporations, investors, and security markets gets disseminated through the Internet almost instantaneously. For the most part, the available information is unstructured in the form of a text. It is easy to see that a strategy of trading on information acquired from free form text would become more profitable the faster you are able to read the text. Hence it is not surprising that text analytics is becoming increasingly important on Wall Street.[1] Hoping to capture the current mood of investors, some traders are using computer programs to monitor and decode the words, opinions, rants and even keyboard-generated smiley faces posted on social networking sites.[2] Academia has followed suit. Computerized decoding of "textual information" into quantitative metrics has become an important area of research in financial economics.

This paper is meant to be a teaser to researchers in financial economics that lowers the costs of entry into the field of text analytics. The paper develops, presents and explains a set of simple Perl programs that will allow access to the electronic filing system (EDGAR) used by the U.S. Securities and Exchange Commission (SEC) to disseminate business relevant information. To illustrate how to download and extract information from EDGAR, we use Form 8-K to analyze executive turnover (new hires and departures of corporate executives). We investigate if there is a calendar effect in executive turnover (there is). But the findings on this particular question are not the main point of this paper. Our key contribution is to show how easy it is to access and analyze the various forms that companies and investors file electronically with the SEC.

The empirical literature that uses textual data as their main data source is growing. García and Norli (in press), Phillips and Hoberg (2010), and Kogan et al. (2009) analyze the annual report filed by firms on Form 10-K. Another strand of the literature has focused on textual analysis of newspaper articles: Tetlock (2007) picks up investor sentiment by analyzing newspaper articles on the stock market, while Dougal et al. (2012) use exogenous scheduling of Wall Street Journal columnists to identify a causal relation between financial reporting and stock market performance. Engelberg (2008) analyze earnings announcements, Hoberg and Hanley (2012) study IPO prospectuses.[3]

* Corresponding author at: 4409 McColl, CB #3490, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3490, United States. Tel.: +1 919 962 8404; fax: +1 919 962 2068.
E-mail addresses: diego_garcia@unc.edu (D. García), oyvind.norli@bi.no (Ø. Norli).

[1] Text analytics covers tagging and annotations, word counting, pattern recognition, etc. The purpose of text analytics is to turn an unstructured text into data that can be analyzed in a quantitative fashion.

[2] USA Today, May 4th, 2011, "Wall Street traders mine tweets to gain a trading edge."

[3] Other related papers include Chan (2003), Barber and Odean (2008), Engelberg and Parsons (2011), DellaVigna and Pollet (2009), Loughran and McDonald (2011),

2173-1268/$ – see front matter © 2012 Asociación Española de Finanzas. Published by Elsevier España, S.L. All rights reserved.
http://dx.doi.org/10.1016/j.srfe.2012.04.001

The rest of the paper proceeds as follows. In Section 2 we present some details on the EDGAR filing system. The next section presents a few simple algorithms to extract basic information from 8-K statements filed with the SEC through EDGAR. Section 4 presents an analysis of the calendar effects around aggregate filings of 8-K statements that discuss executive turnover. The last section concludes.

2. EDGAR

Companies and others are required by law to file a number of different forms with the U.S. Securities and Exchange Commission (SEC). The main purpose of filing these forms is to make certain types of information available to investors and corporations – and by that improve the efficiency of security markets. Historically these forms have been filed with the SEC on paper. In the early 1990s the SEC developed the Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system to handle electronic form filing.[4] As of May 6, 1996 all public U.S. companies were required to make all their filings, with a few exceptions, on EDGAR. More importantly, any person with access to a computer linked to the Internet can obtain and read these filings within seconds after they are filed.

As researchers looking for relevant information on companies with operations in the United States, we have traditionally relied on databases including Compustat, ExecuComp, SDC Platinum, etc. These databases are attractive because their owners have collected data from companies' filings and organized the information in a structured way. Most of the information that is found in Compustat comes from Form 10-K. Most of the information in ExecuComp comes from Proxy filings (Form DEF 14A) and Forms 3–5. The merger information in SDC Platinum relies heavily on the forms filed during the period leading up to a merger. Since the introduction of EDGAR, researchers have had easy access to this "standard" information in addition to an enormous amount of information not found in any other database.

To get an idea of what type of information is available through EDGAR, we move on to looking at the most common forms filed with the SEC. Table 1 reports the filing frequency of the 20 most commonly filed forms over the period 1994–2011. The first column in the table contains a short description of the form. The second column contains the form code used on EDGAR. The third column contains the total number of times a form is filed during the whole sample period.

The most common EDGAR filing is Form 4. For the sample period 1994–2011 this form is filed more than four million times. Form 4 is used to report purchases or sales of securities by persons who are the beneficial owner of more than 10 percent of any class of any equity security, or who are directors or officers of the issuer of the security. This form would, for example, allow you to study the granting of options to officers or directors. Table 1 also shows that Form 4/A is a commonly used form. When "/A" is appended to a form code it means that the filing is an amendment to an existing filing. Thus, a specific corporate event could be linked to an initial filing and a subsequent string of amendments to this initial filing. Form 3 and Form 5, also prevalent in EDGAR, deal with similar ownership issues.

Table 1
Most frequent EDGAR filing codes.

Form	Form code	Frequency
Changes in ownership	4	4,028,202
Current report filing	8-K	1,030,605
5% passive ownership triggers amendments	SC 13G/A	433,902
Quarterly report	10-Q	422,366
Initial ownership report	3	391,533
Definite materials	497	286,299
5% passive ownership triggers	SC 13G	278,735
Current report of foreign issuer	6-K	230,052
Change on a prospectus	424B3	188,880
5% active ownership triggers amendments	SC 13D/A	165,010
Changes in ownership amendments	4/A	150,442
Annual report on ownership changes	5	149,358
Annual report	10-K	128,566
Regulation D exemption, issuance	REGDEX	126,754
Quarterly holdings, institutional managers	13F-HR	126,016
Proxy statements	DEF 14A	125,634
Quarterly report, small business	10QSB	120,120
Registration of securities, investment companies	24F-2NT	116,126
Registration management investment companies	485BPOS	112,269
5% active ownership triggers	SC 13D	86,722

The table presents the frequencies of appearances of different types of filings in the EDGAR database. The time period is 1993–2011. A filing is considered if and only if the same text string (i.e. "4") appears in the form field of the EDGAR master files.

The second most common EDGAR filing is Form 8-K, with more than one million filings. Companies have to use this form to file information on issues that are of "material importance" for the firm. The 8-K statements include information on changes in management, new significant contracts, merger negotiations, lawsuits, etc. In the next sections of the paper we will use the 8-K Forms to illustrate how one can use simple computerized parsing to extract information from the EDGAR filings.

Another important subset of EDGAR is comprised of Form SC 13D (commonly referred to as Schedule 13D) and Form SC 13G. Filing of these forms are triggered when someone crosses the 5% ownership threshold in a firm. The 13Ds are "active" investors, say those seeking control of the firm, whereas the 13Gs are from "passive" investors. There are on the order of 1 million such filings (including amendments).

Annually and quarterly statements also figure prominently in the EDGAR system. There are well over 400,000 10-Q forms, and over 100,000 10-K statements. Other forms that come up in the "top-twenty" list in Table 1 are: foreign firms' current reports, Form 6-K; forms having to do with issuance of securities, from prospectuses, such as Form 424B3, to exemptions from regulation D; forms specific to institutional managers, such as quarterly holdings reported on Form 13F.

Filers in the EDGAR system are uniquely identified using the Central Index Key (CIK). For the sample period 1994–2011, there are 452,830 unique CIKs in the EDGAR database. Only a fraction of these CIKs are publicly traded firms. There are many filers that are private firms. These private firms include manufacturing firms, but also hedge funds and mutual funds. You will also receive a CIK if you are filing on behalf of yourself as an individual.

Table 2 reports the number of filers (unique CIKs) that file a particular type of form. We see there are more than 171,000 filers that have filed a form 4 (or an associated amendment) at some point during the sample period. There are about 40,000 filers that have filed 13-Ds, with a very similar number of 13-G filers. The total number of firms that file some type of 10-K report adds up to over 36,000. This is similar in magnitude to the number of firms that file 8-K statements.

García (in press), Solomon (2012), Davis et al. (2007), Loughran and McDonald (2008), Tetlock et al. (2008).

[4] The SEC describes EDGAR as follows: "EDGAR, the Electronic Data Gathering, Analysis, and Retrieval system, performs automated collection, validation, indexing, acceptance, and forwarding of submissions by companies and others who are required by law to file forms with the U.S. Securities and Exchange Commission (SEC). Its primary purpose is to increase the efficiency and fairness of the securities market for the benefit of investors, corporations, and the economy by accelerating the receipt, acceptance, dissemination, and analysis of time-sensitive corporate information filed with the agency." For more information on EDGAR, visit: http://www.sec.gov/edgar.shtml.