edgar: An R package for the U.S. SEC EDGAR retrieval and parsing of corporate filings

Gunratan Lonare, a,∗, Bharat Patil, b Nilesh Raut, c

a Department of Finance, University of North Carolina, Charlotte, NC 28223, USA
b Department of Finance, Syracuse University, Syracuse, NY 13244, USA
c Department of Health Policy, London School of Economics and Political Science, London WC2A 2AE, UK

ABSTRACT

This paper introduces the R package edgar to download and analyze the Securities and Exchange Commission’s (SEC) mandatory public disclosures in the United States. Corporations in the U.S. submit their periodic reports, registration statements, and financial reports electronically to the SEC. The SEC makes these reports publicly accessible to everyone through the Electronic Data Gathering, Analysis, and Retrieval System (EDGAR). As financial reporting is one of the most crucial aspects of the financial system, efficient retrieval of EDGAR filings becomes imperative for analysts and researchers. We summarize the implementation of edgar package that facilitates downloading, parsing, searching, and sentiment analysis of corporate reports.

© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Motivation and significance

It is noteworthy that the U.S. SEC receives the terabytes of mandatory operation and financial statements, popularly called as filings, every quarter from both public and private firms in the U.S.¹ To download and analyze fundamentals, accounting statements, and future growth possibilities of these firms efficiently are vital. Thus, the edgar package [10] developed in R statistical programming software allows researchers, practitioners, and investors to access valuable information from the SEC filings on a large scale.

The SEC's public repository known as the EDGAR system, started in 1993, maintains filings for an individual company, mutual fund, and exchange-traded fund (ETF). This platform allows public use of these filings for research, investment, and analysis purposes. However, the EDGAR web interface allows accessing only one filing at a time. To make systematic decisions, edgar package helps researchers and analysts to retrieve and parse the required information from these filings in bulk, and performs sentiment analyses.

The SEC has improved its server security, including a significant change in its web interface in 2017. Most of the previously developed packages lack interface to the upgraded EDGAR repository. For example, the links to EDGAR server directories mentioned in [8,9] no longer work. Additionally, these packages do not provide access to multiple files in a single query. For example, xbrlDoAll function from XBRL R package only works on

https://doi.org/10.1016/j.softx.2021.100865
2352-7110/© 2021 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Table 1: Comparative analysis of platforms available for EDGAR mining.

| Tool          | Platform | Specifications | Limitations                                      |
|---------------|----------|----------------|--------------------------------------------------|
| edgarWebR [1] | R        | Provides an interface to access the SEC’s EDGAR system | -Lacks bulk mining functionality  
-Only provides the metadata and company information  
-Lacks parsing of filing for important information  
-Lacks local storage management |
| OpenEDGAR [2] | Python   | Downloads filings and offers a filing search tool        | -Lacks functionality for parsing of important information from filings, such as filing header, 8-K events, business description  
-Uses outside relational database  
-Needs computing knowledge to install and use |
| corpcrawl [3] | Python   | Extracts company and its subsidiaries names from 10-K forms | -No storage structure  
-Package is in beta stage  
-No help document |
| python-edgar [4] | Python | Downloads daily index file | -Provides Minimal functionalities  
-Restriction on form types  
-Lacks proper error handling |
| SECEdgar [5]  | Python   | Downloads 10-K, 10-Q, 8-K, and 13-F forms | |}

| Tool          | Platform | Specifications | Limitations                                      |
|---------------|----------|----------------|--------------------------------------------------|
| finreportr [6] | R        | -Provides filing information of a company  
-Extracts financial reports from XBRL annual reports | |}
| XBRL [7]      | SAS      | Extracts financial statements from XBRL | -Uses licensed software  
-Restriction on form type  
-No storage structure |
| [8]           |          | Downloads filings and provides keyword search function | -Requires large operating time  
-Redundancy in script  
-Inefficient storage structure  
-Restrictions on filing types |
| [9]           | Perl     | -Downloads quarterly indexes  
-Downloads and analyses 8-K filings | |}

Table 1 reports major limitations of the similar tools that provide functionalities to download and parse EDGAR filings. The open-source edgar R package mitigates these limitations and adds new routines. Specifically, it provides an open source tool, access to all filings, robust error handling, better file management system, and scraping and parsing functions. The recent developments in the advancement of research methodologies led to an increase in demand for the usage of such specialized tools.

2. Software description

The edgar package utilizes functions from R.utils [11], tm [12], XML [13], stringr [14], stringi [15], and qdapRegex [16] R packages. A user can install it in R using the following code:

```r
install.packages(“edgar”)
```

2.1. Software functionalities

Table 2 reports all the functions provided by edgar package.

2.2. Software architecture

Efficient download and analysis of a large number of filings require proper storage management. edgar package uses a working directory on a user’s machine to store data in a hierarchy structure. It automatically creates all the sub-directories in the selected working directory upon respective function calls. As seen in Fig. 1, other functions in edgar R package call getMasterIndex and getFilings functions to retrieve filing information and download filings from the SEC server, respectively. We recommend, though not mandatory, to maintain the same working directory for every interaction with this package to utilize the existing data.

edgar package stores filing information, complete filings, and extracted data in separate sub-directories illustrated as follows.

- **Daily Indexes**: This directory is generated upon calling the getDailyMaster function and contains daily filing information, also known as daily master index files, in Rda format with filename as daily_idx_[index date].Rda.

- **Master Indexes**: This directory is created upon calling the getMasterIndex function and saves yearly master index in Rda format with filename as [Year]master.Rda, e.g., 1994master.Rda.

- **Edgar filings_full text**: This directory is generated when a user calls the getFilings function. It stores complete filings in text format in separate sub-directories of form types and firm CIK number with filename as [CIK]_[form type]_[date filed]_[Accession Number]. For example, 10-K statement of a firm with CIK = 38079 can be found at the location “Edgar filings_full text/form 10-K/38079_10-K_2005-03-15_0001047469-05-006546.txt”.

- **Edgar filings_HTML view**: This directory is created upon a call of the getFilingsHTML function and saves filings in HTML format with the filename in format [CIK]_[form type]_[date filed]_[Accession Number]. These HTML files are stored in separate sub-directories of form types and firm CIK number. For example, HTML view of 10-K statement in the previous example can be found on file path “Edgar filings_HTML view/form 10-K/38079_10-K_2005-03-15_0001047469-05-006546.html”.

- **Keyword search results**: This directory is created upon use of the searchFilings function and saves the extracted filing search results in HTML format. The HTML view of search results provides the extracted filing text surrounded by the user keywords. These results for each filing is stored in the format of [CIK]_[form type]_[date filed]_[Accession Number].

---

2. **Rda** is a native R data structure to store objects of vectors, matrices, and dataframes.
3. The central index key (CIK) acts as a primary identifier for corporations and individuals who file disclosures with the US SEC.
Table 2
Functions provided by edgar package.

| Function          | Description                                                                 |
|-------------------|-----------------------------------------------------------------------------|
| getDailyMaster    | Retrieves daily master index                                                |
| getMasterIndex    | Retrieves quarterly master index                                            |
| getFilings        | Retrieves filing information of a firm                                      |
| getFilingsHTML    | Get HTML view of EDGAR filings                                              |
| getFilingHeader   | Scrapes EDGAR filing header information                                      |
| searchFilings     | Searches EDGAR filings for specific words                                   |
| getBusinDescr     | Retrieves business descriptions from annual reports                         |
| getMgmtDisc       | Retrieves MD&A section from annual reports                                 |
| get8KItems        | Retrieves Form 8-K event information                                        |
| getSentiment      | Provides sentiment measures of EDGAR filings                                |

Fig. 1. Architecture of edgar R package.

- **Business descriptions text**: This directory stores extracted business description section in text format with filename as [CIK]_[form type]_[date filed]_[Accession Number]. It is created with usage of the `getBusinDescr` function.
- **MD&A section text**: This directory saves extracted Management’s Discussion and Analysis (MD&A) section in text format with filename as [CIK]_[form type]_[date filed]_[Accession Number]. It is created upon a call of the `getMgmtDisc` function.

3. Implementation of edgar package

3.1. SEC guidelines on downloading EDGAR files

To avoid load congestion on EDGAR server, the SEC suggests downloading only the required filings. The SEC makes it mandatory for a user to declare user-agent id in the request header. The given link explains this requirement in detail: [https://www.sec.gov/os/accessing-edgar-data](https://www.sec.gov/os/accessing-edgar-data). Accordingly, edgar package requires a user to pass user-agent id for all its functions. Thus, a user needs to declare it in the following form:

```r
R> useragent = "Your Name Contact@domain.com"
```

3.2. Download filing information from the SEC EDGAR server

The U.S. SEC receives financial reports regularly from various public and institutional firms. The SEC generates quarterly index files (also called as master index) with the information on all the filings filed on the SEC in a given quarter. The quarterly master indexes are uploaded on the SEC server in idx (index) compressed formats on [www.sec.gov/Archives/edgar/full-index/](https://www.sec.gov/Archives/edgar/full-index/), which includes the Central Index Key (CIK) number, company name, form type, date filed, and weblink for financial reports. The `getMasterIndex` function of edgar package downloads these quarterly master indexes by taking a vector of years as a user input. This function downloads quarterly master index files, cleans them, consolidates quarterly indexes to yearly, and stores them as yearly master index files in `Rda` format in the directory “Master Indexes”. A user needs to maintain the same working directory while using edgar packages as it utilizes these yearly master indexes to search for filing information and download filings from the EDGAR server. The following code illustrates a use of this function.

```r
R> getMasterIndex(2006, useragent)
```

Similar to the quarterly index files, the SEC also maintains index files for filings filed on the SEC in a given day. These daily information on filing can be obtained using `getDailyMaster` function. Thus function takes a date as an input from a user, downloads and cleans the daily index file, and returns information on daily filings in a dataframe.
3.3. Search for filing information and download filings

The getFilingInfo function provides filing information of a firm based on a firm identifier. It takes a desired firm identifier in the form of full/partial firm name or CIK number, filing year(s), filing quarter(s), and form type(s) as input parameters. The following code demonstrates the usage of this function.

```r
R> info <- getFilingInfo('United Technologies',
  + filing.year = c(2005, 2006), quarter = c(1, 2),
  + form.type=c('10-K','DEF 14A'), useragent)
```

The yearly master index files generated using getMasterIndex function contain filing information along with partial links for the complete filings uploaded on the SEC's EDGAR server. The getFiling function facilitates downloading of filings by taking CIK(s), form type(s), filing year(s), and filing quarter(s) as function parameters. The following is an example for implementing this function.

```r
R> output <- getFilings(cik.no = c(1000180, 38079),
  + form.type = c('10-K', '10-K405', '10KSB', '10KSB40'),
  + filing.year = c(2005, 2006), userlist, useragent)
```

The getFilings function downloads complete submission filings, which are in text format, from the SEC server. It uses firm CIK(s) and filing year(s) as input parameters. It downloads the required 8-K filings from SEC, and the extraction status, with the value of one being successfully extracted. The following functions also return a dataframe with filing information and the extraction status, with the value of one being successfully extracted.

```r
R> output <- get8KItems(cik.no = c('1000180', '38079'),
  + filing.year = c(2005, 2006), useragent)
```

3.4. Extract filing header information and search filings for input keywords

Analysts may need filing header information for a firm, such as the period of the report, SIC code, business address. The getFilingHeader function takes an input of CIK(s), form type(s), and filing year(s), and scrapes header information of the required filings. The following code illustrates its usage.

```r
R> header.df <- getFilingHeader(cik.no = c('1000180', '38079'),
  + form.type = '10-K', filing.year = 2006,
  + useragent)
```

Researchers often use qualitative information on financial reports. Especially a plethora of studies use count of specific keywords mentioned in financial reports to develop a qualitative proxy. The getSearchFilings function that searches filings for a user keyword list and returns the count of its mentions (nword.hits) along with filing information. A user needs to provide a search keyword list along with CIK(s), form type(s), and filing year(s). The following code demonstrates the use of this function.

```r
R> output <- searchFilings(cik.no = c('1000180', '38079'),
  + word.list = c('foreign exchange exposures', 'currency transactions'),
  + filing.year = c(2005, 2006), userlist, useragent)
```

The searchFilings function also generates detailed search result for each filing in the directory “Keyword search results”, in HTML format. With this search results, a user can see exact position of the input words in the filing and other surrounding text of at most 250 characters. For example, the generated file “Keyword search results--1000180_10-K_2005-03-18_0000950134-05-005462.html” from the previous command shows the search result.

```r
R> output <- searchFilings(cik.no = c('1000180', '38079'),
  + word.list = c('foreign exchange exposures', 'currency transactions'),
  + filing.year = c(2005, 2006), userlist, useragent)
```

3.5. Extract business description and MD&A sections from annual statements

In recent years, research using the textual analysis of firms’ product/business description section and Management’s Discussion and Analysis (MD&A) section in 10-K has witnessed an exponential increase [e.g.,17–19].

The getBusinDescr function in edgar package facilitates analysts and researchers to extract business description (MD&A) information for desired firms in a single command. It uses firm CIK(s) and filing year(s) as input parameters. It sequentially reads 10-K filings, removes HTML tags, extracts business description (MD&A) sections, and stores them in text files in the directory “Business descriptions text” (“MD&A section text”). These functions also return a dataframe with filing information and the extraction status, with the value of one being successfully extracted.

```r
R> output <- getBusinDescr(cik.no = c('1000180', '38079'),
  + filing.year = 2005, useragent)
```

3.6. Retrieve form 8-K items information

The get8KItems function provides a tool to extract Form 8-K events. This function takes firm CIK(s) and filing year(s) as input parameters. It downloads the required 8-K filings from SEC, reads them sequentially, and extracts event information using regular expressions. The output dataframe contains Form 8-K events information along with filing information. The following code illustrates the use of this function.

```r
R> output <- get8KItems(cik.no = c('1000180', '38079'),
  + filing.year = c(2005, 2006), useragent)
```
3.7. Generate sentiment measures of SEC filings

The `getSentiment` function provides sentiment measures of SEC filings. It takes the help of Loughran–McDonald (LM) sentiment dictionaries [20] to compute sentiment measures of the filing text. The following code illustrates its usage.

```R
> senti.df <- getSentiment(cik.no = c('1000180', '38079'),
  + form.type = c('10-K', '10-Q'),
  + filing.year = 2006, useragent)
```

4. Impact

The popularity of using financial statements filed on the SEC has exponentially increased since the last decade. The usage of these corporate filings is growing in multitude of areas, such as law, accounting, finance, marketing, management, statistics, environment science. However, SEC’s web server provides a single filing at a time. This creates a need of automating download of these filings in bulk with an ease. Additionally, the growing topics using text-mining of SEC filings call attention to develop a tool that helps analysts and researchers for preprocessing of these filings. To fulfill this gap, edgar R package provides functions for downloading, parsing, searching, and sentiment analysis of filings.

5. Conclusions

Post-2000 era has seen an unprecedented rise in the textual analyses research, using financial and operational disclosure of firms, leading to an increased demand for an efficient open-source platform to download and analyze the disclosures. This paper introduces edgar R package in details and could serves as a primer for researchers, practitioners, and investors alike to achieve their respective goals using SEC EDGAR filings. This package works on major operating systems with greater simplicity, providing 11 functions to facilitate retrieving, storing, searching, and parsing of all the available filings on the SEC’s EDGAR server. The edgar package have been undergone several updatations since its implantation in 2015. We plan to extend its features by adding functions to scrape other important information from filings based on the future requirements by the academicians and analysts.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Waldstein M. EdgarWebR: SEC filings access. 2021, URL https://mwaldstein.github.io/edgarWebR, R package version 1.1.0.
[2] Bommarito MJ, Katz DM, Detterman E. OpenEDGAR: Open source software for SEC EDGAR analysis. 2018, Available At SSRN 3194754 URL doi:10.2139/ssrn.3194754.
[3] Rozap. Corpcrawl 0.0.37: Python scraper for the securities and exchange commission EDGAR, 2013, URL https://www.cnpython.com/pypi/corpcrawl.
[4] Eduard S. Python-edgar 3.0.1: Download the SEC EDGAR index since 1993. 2014, URL https://pypi.org/project/python-edgar.
[5] Rahul R. SECEdgar 0.1.1: Implements a basic sphinx crawler for downloading the filings. 2014, URL https://github.com/rahulrrxse/SEC-
edgar.
[6] Lee S. Finreportr: Financial data from U.S. securities and exchange commission. 2016, URL https://CRAN.R-project.org/package=finreportr, R package version 1.0.1.
[7] Bertolusso R. XBRL: Extraction of business financial information from 'xbrl' documents. 2017, URL https://CRAN.R-project.org/package=XBRL, R package version 0.9.18.
[8] Engelberg J, Sankaraguruswamy S. How to gather data using a web crawler: An application using SAS to search EDGAR. 2007, Available At SSRN 1015021 URL doi:10.2139/ssrn.1015021.
[9] García D, Noril Ø. Crawling edgar. Span Rev Final Econ 2012;10(1):1–10, URL https://doi.org/10.1016/j.srfe.2012.04.001.
[10] Lonare G, Patil B. Edgar: Tool for the U.S. SEC EDGAR retrieval and parsing of corporate filings. 2021, URL https://cran.r-project.org/web/packages/edgar, R package version 2.0.4.
[11] Bengtsson H. RUtils: Various programming utilities. 2019, URL https://CRAN.R-project.org/package=RUtils, R package version 2.9.2.
[12] Feinerer I, Hornik K. Txr: Text mining package. 2019, URL https://CRAN.R-project.org/package=tm, R package version 0.7-7.
[13] Temple Lang D. XML: Tools for parsing and generating XML within R and S-Plus. 2020, URL https://CRAN.R-project.org/package=XML, R package version 3.99-0.3.
[14] Wickham H. Stringr: Simple, consistent wrappers for common string operations. 2019, URL https://CRAN.R-project.org/package=stringr, R package version 1.4.0.
[15] Gagolewski M, Tartanus B, other contributors; IBM and Unicode and Inc. and other contributors; Unicode and Inc., stringr: Character string processing facilities. 2020, URL https://CRAN.R-project.org/package=stringr, R package version 1.4.6.
[16] Rinker T. qdapRegex: Regular expression removal, extraction, and replacement tools. 2017, URL https://CRAN.R-project.org/package=qdapRegex, R package version 0.7.2.
[17] Hoberg G, Phillips G. Product market synergies and competition in mergers and acquisitions: A text-based analysis. Rev Financ Stud 2010;23(10):3773–811, URL https://doi.org/10.1093/rfs/hhp053.
[18] Hoberg G, Phillips GM. Text-based industry momentum. J Financ Quant Anal 2018;53(6):2555–88, URL https://doi.org/10.1017/S0022109018000479.
[19] Loughran T, McDonald B. Textual analysis in accounting and finance: A survey. J Account Res 2016;54(4):1187–203, URL https://doi.org/10.1111/j.1475-679X.2012.12123.
[20] Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. J Finance 2011;66(1):35–65, URL https://doi.org/10.1111/j.1540-6261.2010.01625.x.