Analysis of Companies' Non-financial Disclosures: Ontology Learning by Topic Modeling

Andy Moniz[1(✉)] and Franciska de Jong[1,2]

[1] Erasmus Studio, Erasmus University, Rotterdam, The Netherlands
moniz@rsm.nl, f.m.g.dejong@eshcc.eur.nl,
f.m.g.dejong@utwente.nl
[2] Human Media Interaction, University of Twente, Enschede, The Netherlands

Abstract. Prior studies highlight the merits of integrating Linked Data to aid investors' analyses of company financial disclosures. Non-financial disclosures, including reporting on a company's environmental footprint (*corporate sustainability*), remains an unexplored area of research. One reason cited by investors is the need for earth science knowledge to interpret such disclosures. To address this challenge, we propose an automated system which employs Latent Dirichlet Allocation (LDA) for the discovery of earth science topics in corporate sustainability text. The LDA model is seeded with a vocabulary generated by terms retrieved via a SPARQL endpoint. The terms are seeded as lexical priors into the LDA model. An ensemble tree combines the resulting topic probabilities and classifies the quality of sustainability disclosures using domain expert ratings published by Google Finance. From an applications stance, our results may be of interest to investors seeking to integrate corporate sustainability considerations into their investment decisions.

Keywords: Automated ontology learning · Topic modeling · LDA · Sustainability

1 Introduction

Prior studies [1, 2] highlight the benefits of employing Linked Data for investment analysis, by combining information from Dbpedia, stockmarket patterns and different taxonomy versions of companies' accounting statements. Increasingly, investors and regulators are demanding companies to disclose non-financial information, particularly firms' impacts on the environment (referred to as *sustainability*) [4]. The voluntary nature of corporate sustainability reporting has resulted in the publication of inconsistent and incomplete information [4]. This has inhibited the manual creation of ontologies [3, 8]. In this study, we employ an automated ontology learning system to overcome this challenge. The proposed system, labelled SPARQL LDA, employs Latent Dirichlet Allocation (LDA) [5] for the discovery of topics to represent ontology concepts [6–8].

The system works in three phases. The first phase employs a Naïve Bayesian model to categorize text in sustainability disclosures. The model detects text related to a firm's

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9341, pp. 97–101, 2015.
DOI: 10.1007/978-3-319-25639-9_19

climate change impacts and aggregates sentences to create a composite document. The second phase employs a LDA topic model to detect contextual information in text. Topics are learned by retrieving terms via a SPARQL endpoint which are seeded as lexical priors into the LDA model. The final phase combines the LDA topic probabilities in an ensemble model and classifies the quality of corporate sustainability reporting using publically available disclosure ratings.

The rest of this study is structured as follows: Sect. 2 provides a brief overview of relevant sustainability datasets. In Sect. 3 we develop a system to evaluate the quality of corporates' sustainability disclosures. Section 4 provides an empirical evaluation of the proposed system. We conclude in Sect. 5.

2 Environmental Sustainability Datasets

Prior earth science literature has explored the benefits of incorporating Semantic Web technologies to predict the impacts of climate change [9–12]. To our knowledge, literature has not considered the implications for companies or government regulatory policy. To aid such analysis we highlight two publically available datasets. The US Global Change Research Act of 1990 requires a National Climate Assessment (NCA) report [13] on the impact of climate change and affected industries. This includes a Global Change Information System (GCIS) which stores climate change metadata. GCIS resources are exported into a triple store queryable through a public SPARQL interface. A second dataset, published under the "Key stats and ratios" section of Google Finance, provides ratings to evaluate the quality of firms' sustainability disclosures. These ratings are collected by the Carbon Disclosure Project (CDP), an initiative led by the United Nations, and are computed from annual surveys of domain experts. The highest CDP rating, 'A', corresponds to companies that are perceived to have published comprehensive climate change disclosures. The lowest rating, "E", corresponds to companies with poor quality disclosures.

3 Model of Corporate Sustainability

3.1 Climate Change Aspect Detection

The first phase of the system employs a Naïve Bayesian classifier to detect salient aspects in text. A pre-processing step selects classification features from Wikipedia's 'Carbon emissions reporting' page. The page provides an overview of corporate environmental reporting issues. We select the 10 most frequently occurring unigrams and bigrams as classification features: "climate", "climate change", "emissions", "emitters", "gas", "ghg", "greenhouse", "scope 1", "scope 2", "scope 3".

3.2 LDA Topic Model

The second phase of the system employs a LDA model [5] for the discovery of topics represented as ontology concepts [6–8, 16]. In LDA, a topic is modeled as a probability

distribution over a set of words represented by a vocabulary and a document as a probability distribution over a set of topics. Our approach departs from a traditional LDA model [5] by seeding terms as lexical priors following the approach of [14]. Figure 1 displays the SPARQL query which retrieves the key recommendations from the latest NCA report using the GCIS interface (see Sect. 2). The unique terms (excluding stopwords) generated by the query form the LDA model's vocabulary.

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT str($statement) as $statement $finding
FROM <http://data.globalchange.gov>
WHERE { $report dcterms:title "Climate Change Impacts in the United States: The Third National Climate Assessment"^^xsd:string .
  $report gcis:hasChapter $chapter .
  $finding gcis:isFindingOf $chapter .
  $finding dcterms:description $statement . }
```

Fig. 1. SPARQL query used to retrieve a earth science terms

We implement standard settings for LDA hyperparameters with $\alpha = 50/K$ and $\beta = .01$ [15]. The number of topics, K, is set to five following a heuristic approach based on the number of climate change topics reported in the latest NCA report [13]. Table 1 displays the top terms associated with the topic clusters. Cluster labels are manually annotated to aid the reader's interpretation.

Table 1. Topic clusters and top words identified by SPARQL LDA

Footprint	Mitigation	Adaptation	Monitoring	Risks
Emissions	Processes	Systems	Monitoring	Risks
Impacts	Responses	Adaptation	Usage	Regulatory
Ocean	Plans	Goal	Volume	Reporting
Climates	Requirements	Thresholds	Percentile	Policymakers
Ecosystems	Reported	Technology	Stabilizing	Trends
Society	Estimates	Operational	Target	Economic
Reef	Measures	Achieving	Consumption	Shifts
Glacier	Mitigation	Improvements	Percent	Effects
Forest	Research	Target	Capacity	Changing

The outcome of the model is a finer-grained categorization of companies' disclosures based on topics discussed by the online scientific community. The probabilities associated with each topic cluster are included as components within the ensemble tree.

4 Ensemble Model

In this section we outline the evaluation of the ensemble classification tree, present the results and briefly conclude.

4.1 Data

Sustainability disclosures are reported annually on company websites. We retrieve a sample of 443 reports via the Google search query: "sustainability report type:pdf site:" followed by companies' urls obtained from DBpedia (dbpedia-owl:wikiPageExternalLink). Document text is extracted using PDFMiner. To evaluate the ensemble tree's classifications, we create a Boolean which takes a value of one if a company's CDP disclosure is 'A' rated and zero otherwise (see Sect. 2).

4.2 Experimental Setup

We design the evaluation by comparing two systems. The benchmark employs a traditional LDA model and infers topics using only the underlying collection of documents. The SPARQL LDA system incorporates lexical priors by seeding the SPARQL generated vocabulary (see Sect. 3.2). Any differences in classification between the two systems can be explained by the different approaches to topic learning. Experiments were validated using 10-fold cross validation. The performance is evaluated in terms of Precision, Recall, and F1-measure:

$$recall = \frac{TP}{TP + FN} \quad precision = \frac{TP}{TP + FP} \quad F1\,measure = \frac{precision * recall}{precision + recall}$$

The evaluation metrics are shown in Table 2.

Table 2. System evaluation

System	Precision	Recall	F1-measure
Benchmark	0.52	0.59	0.55
SPARQL LDA	0.69	0.65	0.67
%difference	32.7 %	10.2 %	21.8 %

Precision for the SPARQL LDA system improves by 33 % versus the traditional LDA approach.

5 Conclusion

The manual building of ontologies is a time-consuming and costly process particularly in fast evolving domains of knowledge such as earth science, where information is updated often. In this paper we employ a fully-automated method for learning ontologies to alleviate the need for manual approaches. Our findings point to the benefits of integrating Linked Data for investors' analyses of both financial and non-financial disclosures.

Acknowledgement. The research leading to these results has partially been supported by the Dutch national program COMMIT.

References

1. Kämpgen, B., Weller, T., O'Riain, S., Weber, C., Harth, A.: Accepting the XBRL challenge with linked data for financial data integration. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 595–610. Springer, Heidelberg (2014)
2. Carretié, H., Torvisco, B., García, R., Carlos, J.: Using semantic web technologies to facilitate XBRL-based financial data comparability. In: FEOSW (2012)
3. O'Riain, S., Curry, E., Harth, A.: XBRL and open data for global financial ecosystems: a linked data approach. Int. J. Acc. Inf. Syst. **13**, 141–162 (2012)
4. Coburn, J., Cook, J.: Cool Response: The SEC & Corporate Climate Change Reporting. Ceres (2014)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Wong, W., Liu, W., Bennamoun, M.: Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances. IGI Global, Hershey (2011)
7. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag New York, Secaucus (2006)
8. Wei, W., Barnaghi, P., Bargiela, A.: Probabilistic topic models for learning terminological ontologies. IEEE Trans. Knowl. Data Eng. **22**, 1028–1040 (2009)
9. Pouchard, L., Branstetter, M., Cook, R., Devarakonda, R., Green, J., Palanisamy, G.: A linked science investigation: enhancing climate change data discovery with semantic technologies. Earth Sci. Inform. 63, 175–185 (2013) (Oak Ridge National Laboratory)
10. Bozic, B., Peters-Anders, J., Schimak, G.: Ontology mapping in semantic time series processing and climate change prediction. In: 7th International Congress on Environmental Modelling (2014)
11. Emile-Geay, J., Eshleman, J.A.: Toward a semantic web of paleoclimatology. Goechem. Geophys. Geosyst. **14**, 457–469 (2013)
12. Tilmes, C., Fox, P., Ma, X., McGuinness, D.L., Privette, A.P., Smith, A., Waple, A., Zednik, S., Zheng, J.G.: Provenance representation for the national climate assessment in the global change information system. IEEE Trans. Geosci. Remote Sens. **51**, 5160–5168 (2013)
13. Melillo, J.M., Richmond, T.T., Yohe, G.W.: Climate Change Impacts in the United States: The Third National Climate Assessment. U.S. Global Change Research Program, Washington (2014)
14. Jagarlamudi, J., Daume III, H., Udupa, R.: Incorporating lexical priors into topic models. In: EACL (2012)
15. Griffiths, T., Steyvers, M.: A probabilistic approach to semantic representation. In: Conference of the Cognitive Science Society (2002)
16. Zavitsanos, E., Paliouras, G., Vouros, G.A., Petridis, S.: Discovering subsumption hierarchies of ontology concepts from text corpora. In: Proceedings of the International Conference on Web Intelligence (2007)