S&P360: Multidimensional Perspective on Companies from Online Data Sources

Michele Berlingerio[1](✉), Stefano Braghin[1], Francesco Calabrese[1],
Cody Dunne[2], Yiannis Gkoufas[1], Mauro Martino[2],
Jamie Rasmussen[2], and Steven Ross[2]

[1] IBM Research Ireland, Dublin, Ireland
{mberling,stefanob,fcalabre,yiannisg}@ie.ibm.com
[2] IBM Watson, Yorktown Heights, USA
{cdunne,mmartino,jrasmus,steven_ross}@us.ibm.com

1 Introduction

We introduce S&P360, a system to analyse and explore multidimensional, online data related to companies, their financial news, and the social impact of them. Our system combines official and crowd-sourced data sources to offer a broad perspective on the impact of financial newsregarding a set of companies. Our system is based on ABACUS [1], a multidimensional community detection algorithm grouping together nodes sharing communities across different dimensions. ABACUS is able to find both explicit connections that appear as direct links among entities, but also hidden ones coming from indirect interactions between nodes. We enrich structural connections (co-occurrence in Twitter and news articles, hyperlinks in Wikipedia) with latent semantics associated to them by applying NLP techniques such as Latent Dirichlet Allocation (LDA), guiding users in interpreting results. We add a powerful visualization interface enabling users to query the data by company, time, and dimension. Users can browse the results and explore the communities along with their associated semantics. "Evidence" of the structural connections (i.e., the source documents supporting the explicit connections) are shown in the user interface, as well as community

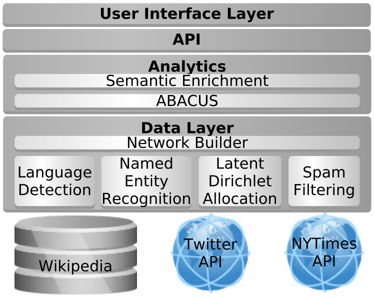


Fig. 1. S&P360 architecture

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 320–324, 2015.
DOI: 10.1007/978-3-319-23461-8_39

topics that summarize both explicit and latent semantic associated to the multidimensional relationships. The challenges faced in the development of S&P360's technology and architecture arise mainly from the noise contained in online data, particularly significant in sources like Twitter, and from the difficulty in displaying and interacting with multidimensional and semantically enriched data effectively. The demo uses publicly available data from Twitter, Wikipedia, and the New York Times (hereafter, NYTimes). We obtained multidimensional communities showing interaction between companies around important events like acquisition of companies or products, change of core business, stock market fluctuations, and the launch of new products. All these events are not hard-coded, but emerge directly from the data and interpretation of the semantics associated to the communities.

2 Architecture

S&P360 is modular, and the architecture of the main modules is depicted in Figure 1.

2.1 Data Layer

This layer implements five main modules: language detection, named entity recognition, latent dirichlet allocation, anti-spam filtering, and network builder.

Language Detection. We used a public library for language detection that is based on a Naive Bayes approach (code.google.com/p/language-detection), for its very good accuracy-scalability trade-off. We focused on English tweets, for the sake of a better manual validation of the next steps, but there are no particular constraints in adding other languages already available in the library.

Named Entity Recognition. After filtering by language, we reduced problematic ambiguous terms like "Apple" or "Coach" by running named entity recognition on the English tweets. We used Stanford's NER (nlp.stanford.edu/software/CRF-NER.shtml) library for its popularity and its specific training for organizations.

Latent Dirichlet Allocation. In order to select tweets relevant to finance, we ran Latent Dirichlet Allocation (LDA) to assign to each tweet a scored set of related topics, in a semi-supervised fashion. We used MALLET (mallet.cs.umass.edu/topics-devel.php) as it provides a well established set of tools and links topics back to source documents.

Anti-Spam Filtering. The next step was filtering out "bots", i.e. automatic Twitter accounts talking about companies. A simple and natural way to do this is to rely on their very limited dictionary, so we just kept the tweets maximizing the ratio $\frac{\#distinct\ words}{\#tweets}$.

Network Builder. We take the results of the four previous modules and build a multidimensional network between the companies found. Links are assumed to be bidirectional by default. For Wikipedia, we link companies if the wiki page of one had the other as an external link. In Twitter, we link companies mentioned together and weight by the number of such tweets. In NYTimes, we link companies if they were mentioned together in an article, or if one of the two was used to tag an article regarding the other one. The number of articles for which this happens is used to weigh the edges.

2.2 Analytics

This layer implements two main modules: ABACUS and semantic enrichment.

ABACUS. ABACUS [1] is an innovative algorithm for community detection in multidimensional networks that introduces a new definition of community, i.e. a group of nodes sharing memberships to the same community in the same dimensions. This property allows ABACUS to relax the requirement for dense multidimensional connections between nodes across dimensions, while finding nodes that are related together but not necessarily directly connected in any of the dimensions [1]. This means that we can find a group of companies that are related to each other regarding a set of topics, although they do not have to be directly connected in Wikipedia, or Twitter, or NYTimes.

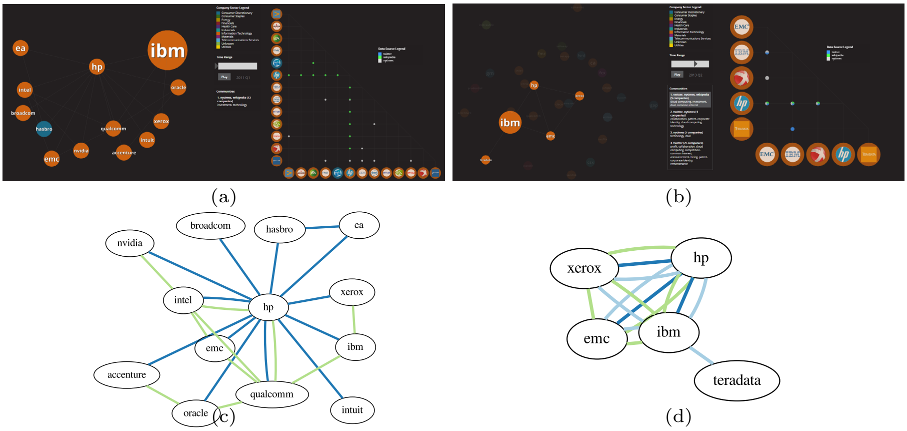
Semantic Enrichment. Two steps are performed in this block: i) for each edge in a resulting community, we retrieve all the associated information including topics with associated score; ii) we compute aggregated topics and scores (min and max across edges) for the community. In this way, we have a semantic associated to the entire community rather than to a single edge, and we can sort or query the communities using the semantics rather than only time or dimension.

2.3 API

S&P360 analytics block exposes a set of API methods to the User Interface Layer. Such calls have been developed as a RESTful Web service, deployed in IBM BlueMix (console.ng.bluemix.net), an implementation of the Cloud Foundry PAAS specification.

2.4 User Interface Layer

S&P360 provides a visual interface to allow interactive exploration of the ABACUS algorithm's results. The user interface is built on HTML5 and JavaScript technologies. Though the visualizations are two-dimensional they are implemented as planar shapes within a 3D scene rendered on WebGL canvases using Three.js (threejs.org), allowing for high rendering performance and smooth zooming. The D3.js (d3js.org) library is used for network layout. Upon user selection of a company and time range, the interface queries the API for up to ten top scoring community matches, which are shown in a ranked list (see Figure 2). Each community result indicates the data source(s), number of companies, and representative topics as determined by semantic enrichment.



(a) (b) (c) (d)

Fig. 2. Examples of resulting communities visualized by S&P360 (top) and with GraphViz (bottom)

Table 1. Characteristics of the two communities. First row is community in Fig 2a,c. Second is Fig 2b,d.

Time	Companies	Dimensions	Topics	Evidences (subset)
Q1 2011	Accenture, Broadcom, EA, EMC, Hasbro, Hp, IBM, Intel, Intuit, Nvidia, Oracle, Xerox	Wikipedia, NYTimes	investment, technology	Wikipedia: IT / electronics companies NYTimes: "Recent Acquisitions by Major Technology Companies http://www.nytimes.com/aponline/2011/01/05/[...]" "[..] Four American giants - Microsoft, Oracle, Amgen and Pfizer - were among the fund's top five [..] The fifth is Accenture" "Computing has shifted to phones, but the leading maker of processors has not. Now Intel is trying to catch up to Qualcomm, Nvidia and Marvell."
Q2 2013	EMC, Hp, IBM, Teradata, Xerox	Wikipedia, NYTimes, Twitter	cloud computing, investment, deal, common interest	Wikipedia: IT / electronics companies NYTimes: "International Business Machines Corp and EMC Corp are among parties in talks to buy privately held database web hosting company SoftLayer Technologies Inc, in a deal that could fetch over $2 billion, three sources close to the matter said." Twitter: "Webinar replay: IBM & Teradata Compared: A Total Cost of Ownership Study http://t.co/JPMUpDHjhU" "RT @mcgoverntheory: Big #Amazon Customer Moves To HP Public #Cloud http://t.co/OaVLS7f6Wt #cio #entarch #ibm #emc #gartner #forrester #ensw.." "Teradata, IBM and EMC are at #hadoopsummit. Oracle and HP are missing."

3 Demo on Real World Data

We tested S&P360 on real data from Wikipedia, Twitter and NYTimes. We started from the list (en.wikipedia.org/wiki/List_of_S%26P_500_companies) of the 500 S&P companies available in Wikipedia, reporting stock ticker symbol, name, sector, address, website, and date entered. We then crawled the landing 500 links from this page and built the Wikipedia network dimension out of it. For sake of simplicity, we ignored the temporal dimension in Wikipedia. This step generated 3,637 edges in the Wikipedia dimension. We used the list of names and tickers to query the NYTimes and Twitter APIs. For Twitter, we had access to 10% of the world wide stream from 2011 to 2014. From this collection, we ran string matching on tickers and names, and ended up with 5,724,590 tweets. Out of these, the language detection took 3,276,367 English tweets. We ran LDA on the resulting dataset using 200 topics. We manually annotated all of them, ending up with 19 topics relevant to the financial space, with 231,075 corresponding tweets. By running the named entity recognition task we selected 79,428 tweets. From those, we filtered out the tweets coming from around 100 automated bots, ending up with a final collection of 55,678 tweets. This step generated 11,966 edges in the Twitter dimension. From NYTimes, we retrieved a total of 103,676 snippets (all in English) by querying company names. We then ran LDA using 100 as number of topics and found 6 topics to be relevant, corresponding to 1,010 articles. This step generated 7,227 edges in the NYTimes dimension. We then ran ABACUS on each (as the networks very small, the running time of ABACUS is less than 2 seconds). ABACUS found 97–158 multidimensional communities (we excluded monodimensional communities). For visualization and interaction, our APIs return only 6–34 communities per quarter. These communities were filtered by number of dimensions (higher first) and minimum topic score. Figure 2 shows S&P360's interface displaying two different results. On the top, we see the two communities in the S&P360 user interface. They were found by querying "IBM" and narrowing the search to Q1 2011 (left) and Q3 2013 (right). Since multidimensional properties, topics, and evidences are available upon user interaction, we report their structure on the bottom of Figure 2 and their characteristics in Table 1.

4 Video and Requirements

A video of S&P360 in action can be seen at vimeo.com/117626196, showing the visual querying interface; the result browsing capability; and the interaction with multiple dimension, topics, and evidences. For best performance, the demo would need a widescreen external monitor and Internet access.

References

1. Berlingerio, M., Pinelli, F., Calabrese, F.: ABACUS: frequent pattern mining-based community discovery in multidimensional networks. Data Min. Know. Dis. **27**(3), 294–320 (2013)