Trucolo and Digiampietri *Journal of the Brazilian Computer Society* (2017) 23:8
DOI 10.1186/s13173-017-0056-9

Journal of the
Brazilian Computer Society

RESEARCH Open Access

Improving trend analysis using social network features

Caio Cesar Trucolo* and Luciano Antonio Digiampietri

Abstract

In recent years, large volumes of data have been massively studied by researchers and organizations. In this context, trend analysis is one of the most important areas. Typically, good prediction results are hard to obtain because of unknown variables that could explain the behaviors of the subject of the problem. This paper goes beyond standard trend identification methods that consider only historical behavior of the objects by including the structure of the information sources, i.e., social network metrics, as an additional dimension to model and predict trends over time. Results from a set of experiments indicate that including such metrics has improved the prediction accuracy. Our experiments considered the publication titles, as recorded in the Brazilian Lattes database, from all the Ph.Ds. in Computer Science registered in the Brazilian Lattes platform for the periods analyzed in order to evaluate the proposed trend prediction approach.

Keywords: Trend analysis, Social network

Introduction

Data-driven activities are getting more and more usual in many types of organizations and data analysis is becoming the main focus of business. In this context, trend analysis is a major application of data analysis. Organizations may try to identify trends to create strategies and plan actions, e.g., an e-commerce company may try to identify trends in order to better focus their supply chain activities.

There are several approaches used for prediction and most of them are based on the temporal behavior of the studied subject. Usually, the temporal behavior is modeled as a time series, where time explains the behavior of the relevant variable. However, when these subjects are produced or consumed by people (journalistic texts or new technology products, for example), another factor can be taken into account: the social structure of the generators or consumers, i.e., individuals directly related to the object under analysis. A social network in this context can be modeled around these individuals. Nodes can represent producers (or consumers) and edges can represent relationships between them. Taking the content of blog posts as an example, a social network can be built based on the connections among bloggers, i.e., the hyperlinks that connect the websites.

The analysis and quantification of the behaviors and relationships of the people in the social structure can be performed using social network analysis. We can calculate social metrics to understand influences, centralities, and communities to predict the information diffusion in the network [10, 15, 17, 18, 24, 25]. As we understand the social network characteristics given the calculated metrics, we become able to identify which individuals will be reached by the information spread. It enables us to say whether it is going to take a long or a short time. For example, an information being propagated by a very influential node within a specific time interval can reach more nodes in the network than if it were propagated by a non-influential network node. The social structure plays an important role in the temporal behavior of objects [8, 24]. This work differs from previous work in that besides using the temporal behavior of the studied object, it incorporates the social structure of the individuals related to this object into the prediction models.

In this paper, we present an approach that combines the prediction models based on the temporal behavior of the studied object with social network metrics. This approach can be applied to improve the accuracy of trend predictions that are based only on the temporal behavior and

*Correspondence: trucolo@gmail.com
University of São Paulo, São Paulo, Brazil

where it is possible to model a social network from the interaction among individuals related to the object. For the purpose of this article, we applied this approach to the academic co-authorship environment. Essentially, we used a corpus of titles of papers published in a certain period to predict what will be the major topics (represented as n-grams) in the future. This problem could be solved with standard trend analysis approaches that rely on predicting future frequencies from the observed ones, whereas in this work we consider the properties of the co-authorship network to enhance the predictions. In this case, the objects considered are the n-grams extracted from the paper titles and the individuals are the paper authors.

The approach was tested and validated using data from the publication titles of Computer Science Ph.D.s working in Brazil and then compared with approaches that consider only the temporal behavior of the analyzed object.

This paper is organized as follows. "Related work" section describes some basic concepts and related work. "Methodology" section details the methodology used. The results are described in "Experiments and results" section. Finally, conclusions are presented in "Conclusion" section.

Related work

Time series trend analysis

Usually time is a very important feature in prediction and classification problems. Once there is an understanding about the object temporal behavior it is possible to identify patterns and predict trends. A problem modeling in which time is considered as an explanatory variable is known as time series analysis [9].

Trend analysis can be applied to several topics, such as stock market [20], textual documents [21], and many others [22]. The trend identification in textual documents, more specifically in a corpus formed by titles of scientific papers, is the application addressed in this paper. In the context of textual documents, frequency counts are usually used as the dependent variable in time series models [1].

Social network trend analysis

There are many ways to model and explore social networks and one of the research branches is trend analysis in social networks. How to measure the dynamism and impact of information flow? To answer this question, it is necessary to study the characteristics of the network and its connection structure, that is, how nodes and edges are distributed in the network. Information is produced and transmitted by individuals and their connection structure affects how information diffuses [3]. A very important characteristic of the individuals in the network is their influence. Finding influential nodes in the network can help to explain how fast the information will spread and how many nodes it will reach. There are methods developed to identify influential nodes [19]. Beyond the individual level, analysis of the size and density of groups in the network is very important to understand the dynamics of information diffusion. For this, it is necessary to identify these groups or communities, which is not a trivial task [16]. Another challenge is to identify critical points in the network where the probability of information diffusion increases [2]. Finally, social network information is being used in several ways to predict trends based on the network behavior [13].

Science and technology systems embrace several utilities related to scholars and knowledge can be discovered in a quantitative way [11]. Research productivity, for example, can be measured by models that use citation indices and academic social network analysis [4]. The application explored in this paper also uses data from a science and technology system aiming to identify research trends and topics.

Our work differs from others by combining time series and social network analysis. The proposed approach uses these two concepts so that trends are identified based on time and social characteristics of the individuals that generate information.

Methodology

The methodology of this work consists of five steps: data gathering, term extraction, time series analysis, social network analysis, and trend analysis. Figure 1 illustrates the schematic data flow. The next sections describe all the steps applied to the problem of trend identification for the publications in Computer Science in the Brazilian academy context. The proposed approach can be applied to improve the accuracy of trend predictions that consider only the temporal behavior in scenarios where it is possible to recover the connection between the individuals who generate the data (e.g., trend identification of topics discussed in the blogosphere).

Data gathering

Brazil maintains a unique platform called Lattes Platform[1]. This is a database of information on science, technology, and innovation, including publications by individual researchers, and currently registers over 4.5 million curricula. In this work, all the information has been obtained from Lattes Platform.

For data gathering, curricula from all the Computer Science PhDs were selected for the periods analyzed (comprising 5642 curricula). The pre-processing consisted of the extraction and organization of the information using the methodology described by Digiampietri et al. [6, 7]. The pre-processing activities include the stop-words removal and coauthorship

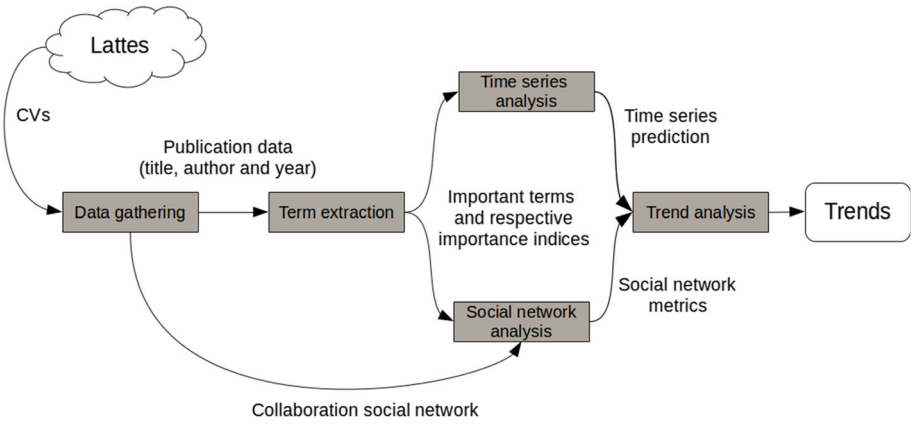

Fig. 1 Schematic data flow of the work process

identification based on an entity resolution approach [7]. From these curricula, 55,710 titles were identified from papers published between the years 1991 and 2012.

The variables considered to build the dataset are lattesId (researcher identification number), year (year of publication), title (title of publication), and publicationId (publication identification).

Term extraction

In this paper, a term is an n-gram extracted from the titles of the papers. In this step, the goal was to automate the data preparation. The first stage of term extraction was to split the titles into subsets of words or sequence of words without *stop-words*. The terms extracted consist of one or more consecutive words from the titles excluding words that were listed as *stop-words*. As an example, the title *Social Network Analysis For Digital Media* was split into the following terms: *Social*, *Network*, *Analysis*, *Digital*, *Media*, *Social Network*, *Network Analysis*, *Digital Media*, and *Social Network Analysis*. Terms such as *Analysis Digital Media* and *Media Digital* are not included because they are not formed by consecutive words from a title or because they include *stop-words*. In this example, we obtained unigrams, bigrams, and 3-grams, however, the process can obtain n-grams for all possible n.

With all the possible sets of terms, we adopted a scoring system to identify the most important terms. This scoring method was based on the adjacent frequency of the words in the terms. The equation to measure the importance of each candidate term is:

$$LRF(CT) = f(CT) \times \left(\prod_{i=1}^{T} (LF(Ni)+1)(RF(Ni)+1) \right)^{1/T} > 1.0.$$

$f(CT)$ is the frequency of the candidate term CT, $LF(Ni)$ and $RF(Ni)$ indicate the frequency of the left and right word candidates, respectively. This equation is described in detail by Nakagawa et al. [12]. In that same work, the authors conducted evaluations to demonstrate that it is possible to find meaningful terms.

In summary, in this step we automatically extract the terms (n-grams) and then filter the most meaningful ones to build our dataset. We observed that n-grams had more significance than the unigrams for the subjects discussed in the publications. Since our goal is to identify terms and research topics, unigrams could be very ambiguous. For example, the word *Network* can be ambiguous given that it can be related to *Social Network*, *Neural Network*, or even *Business Network*. Therefore, we selected the 1638 most important n-grams, this is the number of n-grams occur over all the period (1991–2012) considered in the experiments, as explained in "Experiments and results" section.

Time series analysis

Given a dependent variable and a set of independent ones, a regression model can be formulated as

$$Y \approx f(X, \beta)$$

where the dependent variable Y can be approximated by the independent variables X and the respective parameters β for a function f. For the analysis in this step, we are interested in the frequency (TF-IDF) variation of each term over a target period (e.g., a year). For each term, a time series of its yearly frequency variation is built.

The time series can have many types of shapes and behavior thus we used linear and nonlinear regressions (linear, exponential, logarithmic, power law, and polynomial with 2° to 5°). We applied all for each term and chose the one that best fitted each of the series using ordinary least squares for evaluation. The regression curves for a few terms are shown in Fig. 2.

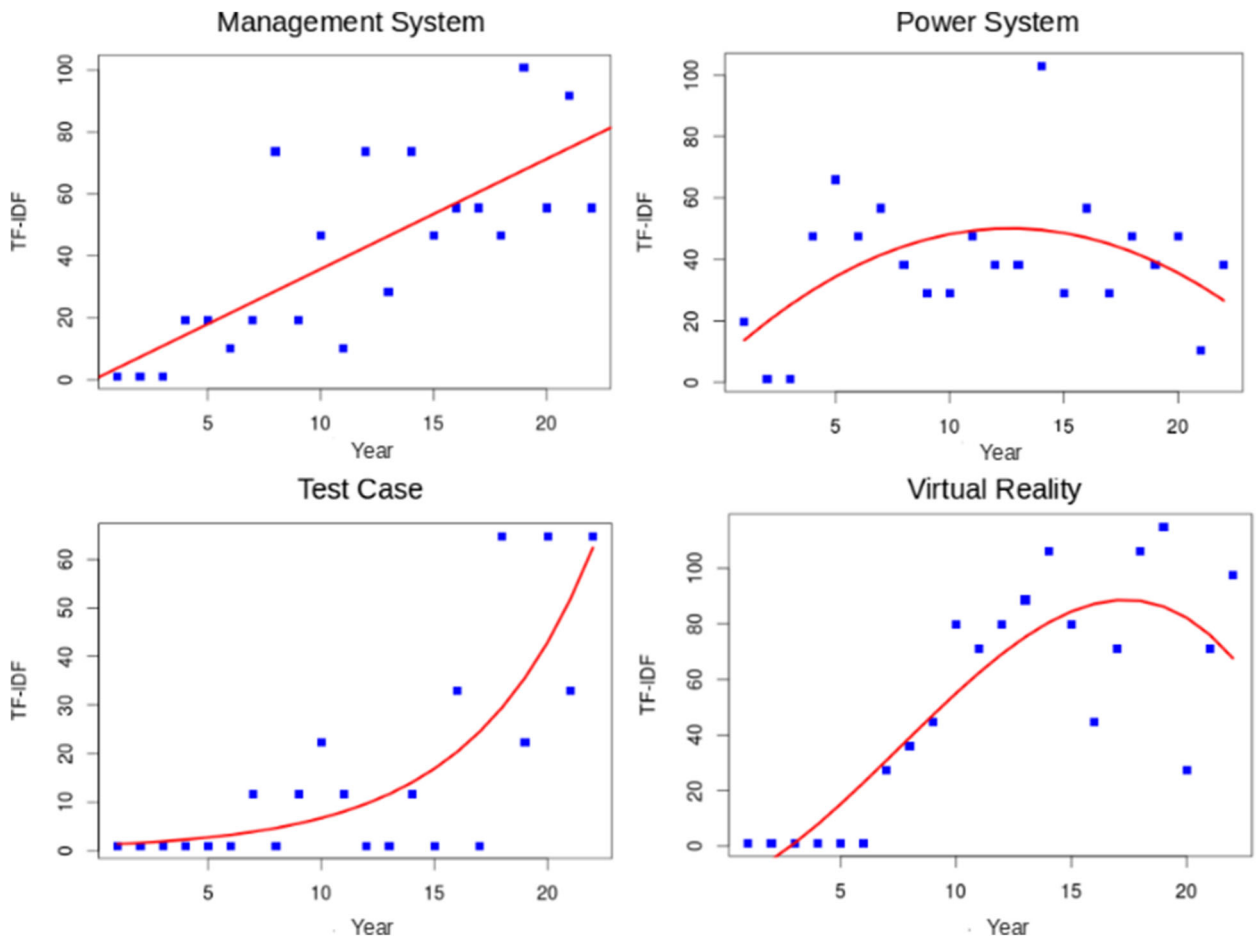

Fig. 2 Examples of regression curves

As a result, we obtained the best prediction among the regression methods cited above for each term to be used for building the datasets for trend analysis. These results are taken as a basis for comparison with the proposed approach.

Social network analysis

The network

The network modeled was built from the joint publications (co-authorship relationships) as recorded in the Lattes database. The social network was modeled as a graph composed of 5642 vertices (authors) and 14,647 edges (coauthorship relationships).

The metrics

Metrics of the social network capture different characteristics that can be quantified. In this approach, some metrics have been selected to form the independent variable's set. Selection was based on assumptions on the potential of each metric to explain the information spreading [14, 23]. For example, one of the assumptions is that a node in the giant component of a network is more capable of disseminating information through the network than a node which is not in this component. The metrics selected are *giant composition, the shortest path to the most central node, degree centrality, eigenvector centrality, page rank centrality, betweenness centrality, closeness centrality, clustering coefficient, structural equivalence to the most central node and community average centrality* [10, 15, 17, 18, 24, 25]. These metrics are described as follows.

Giant composition: number of nodes in the giant component; *Shortest path to the most central node*: smaller value among the shortest paths to the most central node; *Degree centrality*: average degree centrality of the nodes within the community; *Eigenvector centrality*: average eigenvector centrality of the nodes within the community; *Page rank centrality*: average page rank centrality of the nodes within the community; *Betweenness centrality*: average betweenness centrality of the nodes within the

community; *Closeness centrality*: average closeness centrality of the nodes within the community; *Clustering coefficient*: average value of the clustering coefficient from the nodes within the community; *Structural equivalence with the most central node*: average value of the structural equivalence from the nodes within the community; *Community average centrality*: average centrality of all community nodes.

The centrality metrics can explain the importance of a node in the network, the shortest path metric indicates how far a node is from the central node, while the structural equivalence quantifies the similarity of a target node to the most central node. The most important node was used as a reference. To justify this choice, Table 1 shows the difference in the Degree and Eigenvector centralities between the most central node and the other top ten most important nodes in the network.

Each selected metrics has been computed for all network nodes and each term has been related to one or more nodes (a term may have been employed by one or more authors). If a term is related to a single author, then, in this step, its metrics will have exactly the same values of its related node. However if a term is related to multiple nodes then the metrics must be aggregated. The aggregated metrics is computed as the sum of the metrics values of each author related to the term. The one exception is the metrics Shortest path to the most central node, for which the aggregated metrics is taken as the minimum value from all authors related to the term. For example, if *author 1* has a *Degree centrality* of 10 and *author 2* has a *Degree centrality* of 5 and both used a *term A*, the *Degree centrality* value of *term A* would be 15.

We also improved the approach with a so-called network community balance. Communities are characterized as groups of nodes with a high edge density [24]. In a community, information propagates quickly and tends to become general knowledge. In trend analysis, this can lead to situations such as terms that are widespread in a particular community but not in the network as a whole. Thus, it is important to evaluate whether the importance of a term occurs only within a community or in the whole network. Thus, we decided to apply a community level aggregation to balance the node level aggregation in computing the metrics values of each term. To identify the communities, we used the R^2 implementation of the algorithm proposed by Clauset et. al. [5].

Therefore, for nodes that are within the same community, the aggregated metrics value is computed as the average value of the nodes; for nodes that are not in the same community it is computed as the sum of node metrics values. For example, let us assume that a *term A* is used only by two authors: *author 1* and *author 2*. If *author 1* and *author 2* are in the same community, the metrics would be calculated as metrics average of the authors in this community who used *term A*. But if *author 1* and *author 2* are in different communities, the average metrics value would be computed for each community and these results would be finally summed.

Table 1 Eigenvector centrality and number of degrees of top ten central nodes

Top important nodes	Eigenvector centrality	Degree
1	1.000	67
2	0.986	45
3	0.944	45
4	0.845	31
5	0.825	35
6	0.799	29
7	0.798	37
8	0.763	24
9	0.745	30
10	0.744	27

Table 2 Time series and supervised learning methods best results (lowest RAE) for three periods (1991–2011, 2002–2011 and 2007–2011) having 2012 as the year of prediction

Period	Time series regression	Time series supervised learning methods
1991–2011	*113.16%*	51.52%
2002–2011	136.42%	51.01%
2007–2011	288.14%	*50.31%*

The italicized data refers to the best result of each period

Table 3 Basic statistical metrics relative to the final dataset built for the first experiment (period 1991–2011)

Feature	Mean	Std. Dev.	Min	Max
Giant composition	15.18	24.16	0.00	368.00
Shortest path	2.63	1.52	0.00	10.00
Degree centrality	37.04	28.36	0.00	148.73
Eigenvector centrality	0.02	0.03	0.00	0.24
Page rank centrality	0.01	0.01	0.00	0.02
Betweenness centrality	78,492.67	67,884.57	0.00	368,246.75
Closeness centrality	0.01	0.01	0.00	0.07
Clustering coefficient	0.92	0.79	0.00	5.70
Structural equivalence	0.02	0.06	0.00	1.01
Community average centrality	18.93	13.80	0.00	95.21
Time series prediction	14.37	29.96	−97.50	324.46

In the end of this step, the first part of the feature vector is finished which row is a term and each column is a social network metric.

Trend analysis

With the time series analysis and social network analysis performed, we are able to model the behavior of each term. At this moment, the time series model and the social network analysis are combined. Having the social network metrics and the time series prediction calculated, we modeled the problem as the term importance index being explained by the social characteristics and the "clue" about its future importance (TF-IDF predicted by time series prediction methods). The feature vector built in the previous step (as described in the "The metrics" section) is then enriched with the importance index (TF-IDF) predicted by the time series models. Thus, in this step, the dataset to be input into the proposed trend analysis is built.

Both social network analysis and time series analysis rely on the periods considered. Therefore, for each time interval of model training, the dataset will be different. For example, the dataset relative to the period between 2002 and 2005, built to predict 2006, is different from the dataset relative to the period between 2002 and 2006 built to predict 2007, that is, one year (2006) is included in the second case.

The dataset is subject to preprocessing methods such as normalization and feature selection and then supervised learning methods are applied to predict the importance

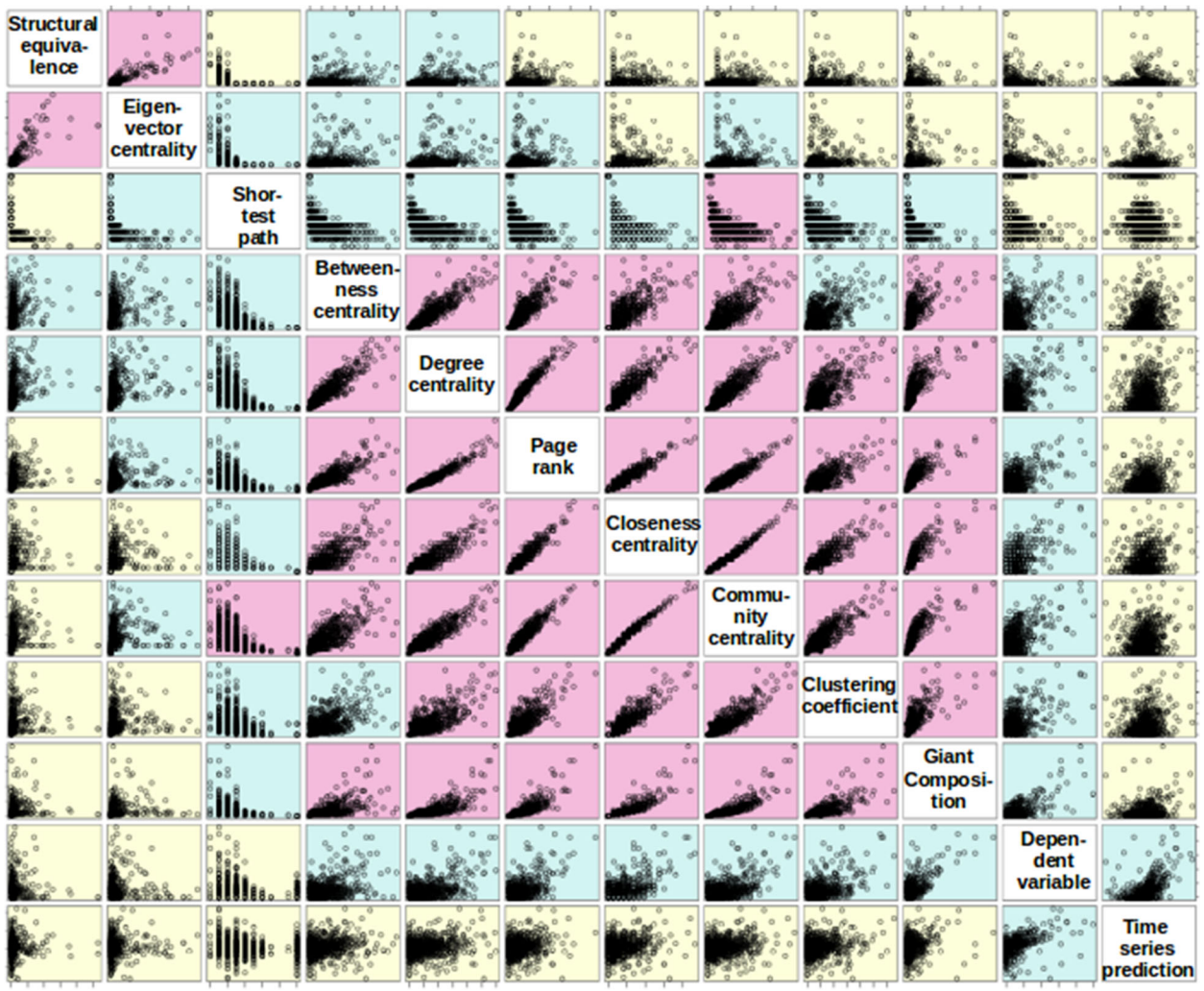

Fig. 3 Scatterplot of each pair of explanatory features and dependent variable (importance index) in the final dataset for the period between 1991 and 2011. The highest correlated pairs of variables are in the center and have the colored *red* while the lowest correlated pairs are towards the border and colored *yellow*

index of the terms for certain periods. The methods considered in the experiments were Linear Regression, Artificial Neural Network (ANN), Support Vector Machine (SVM) and Rotation Forest (RF). In this context, trends are the terms with high predicted values of importance index.

Experiments and results

The main goals of the experiments are to evaluate the proposed approach and compare with results from standard time series prediction. Identifying which models, periods, and variables present a better performance is also within the scope. We split the experiments into two groups. In the first group the goal was to evaluate the best techniques, time period and set of variables while the second group of experiments was designed to evaluate longer prediction periods based on the best set of variables and techniques.

Models were evaluated by measuring the Relative Absolute Error (RAE), comparing the true TF-IDF values observed (y_i) with the predicted ones (f_i). The equation for RAE is

$$RAE = \frac{\sum_{i=1}^{n} |f_i - y_i|}{\sum_{i=1}^{n} |y_i - \bar{y}|}$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

At first, we made experiments based only in the temporal behavior. We considered the same time series dataset model (with no social metric feature, only TF-IDF and year) to evaluate and compare two different kind of techniques: time series methods and supervised learning methods. Table 2 shows the best RAE value relatives to the three periods that represent short, medium and long periods for the time series trend analysis. With the conventional methods for time series analysis, the best result was obtained for the longest period while with the supervised learning methods, the shortest period yielded the best results. We can see that the supervised learning methods were considerably more accurate than the regression ones.

To test the model described in this paper we used the final dataset explained in "Trend analysis" section. A description of this dataset is shown in Table 3.

It is clear that some features take values on a larger scale than others (e.g., Betweenness centrality). To correct these differences we applied normalization in the preprocessing step.

Before applying the prediction methods, a correlation analysis was conducted to clarify the behavior of each feature. Figure 3 depicts the scatterplots showing the pairwise correlations of features, including the dependent variable. We can see that no feature is highly correlated with the dependent variable (importance index) but most of them have some correlation. As expected, most of the centrality metrics are very correlated indicating that some of them can be discarded in the supervised learning step.

In this experiment, we varied the number of features selected. We generated datasets with the instances described by all attributes (features) and datasets with attributes selected by Relief and manual selection, which is an appropriate selection method if the analyst has knowledge about the problem domain. The most important criterion in selecting the features manually was their mutual correlation.

Furthermore, we varied the parameters for each prediction model algorithm generating 16 tests for ANN, 9 tests for SVM, and 15 tests for Rotation Forest. For ANN, we varied the parameters related to learning rate, momentum term, number of nodes in the hidden layer, and number of hidden layers. For SVM, we experimented several kernels (including Radial Basis Function kernel and Polynomial kernel) and different values for parameter C. In Rotation Forest, different tree based methods for the ensemble approach were tested, varying their specific parameters in each case.

Table 4 presents the best RAE results obtained from each model, considering the different feature selection methods, for different periods.

As far as the techniques are concerned, the best performances, as shown in Table 4, were obtained with Rotation Forest. One observes that Rotation Forest achieved the best performances for short periods while SVM did better on longer periods, doing better than Rotation Forest in the 1991–2011 period.

When analyzing the periods, the period 2002–2011 presented the best results considering an average among

Table 4 Best results (lowest RAE) of all prediction methods for three periods and three different feature's sets

Method/Period	All attributes	Manual selection	Relief
Linear Reg. 91–11	72.18%	–	–
ANN 91–11	72.95%	73.43%	73.44%
SVM 91–11	65.21%	*62.21%*	64.15%
Rot. Forest 91–11	71.51%	70.57%	71.18%
Linear Reg. 02–11	53.22%	–	–
RNA 02–11	45.75%	46.25%	46.23%
SVM 02–11	43.45%	41.05%	41.30%
Rot. Forest 02–11	40.29%	*40.04%*	40.12%
Linear Reg. 07–11	62.76%	–	–
RNA 07–11	57.77%	59.02%	57.98%
SVM 07–11	53.31%	52.37%	51.22%
Rot. Forest 07–11	41.68%	*39.28%*	40.33%

The italicized data refers to the best result of each period

Table 5 Sets of features for the model with the best results (lowest RAE)

Relief	Manual selection
Giant component;	Giant component;
Eigenvector centrality;	Eigenvector centrality;
Betweenness centrality;	Clustering coefficient;
Clustering coefficient and	Structural equivalence and
Time series prediction.	Time series prediction.

all techniques, however, the best result was obtained in the 2007–2011 period (39.28%). The average RAE values for the best techniques are: 43.77% for 2002–2011; 51.57% for 2007–2011; and 69.68% for 1991–2011. There is an important difference between the two models at this point. While the time series model yielded better results on longer periods (Table 2), the proposed approach presented better results on shorter periods. This can be explained by a change in the network dynamics. Metrics derived from networks modeling longer periods can be misleading, as network properties are likely to change considerably along time.

Comparing the best results of the proposed approach with the time series model (Tables 2 and 4) one observes an error reduction of 45%, 70% and 86% for the 1991–2011, 2002–2011 and 2007–2011 periods, respectively. While comparing to the supervised methods applied to time series dataset (Tables 2 and 4) one observes an error increase of 21% for the 1991–2011 period and an error reduction of 22% for the 2002–2011 and 2005–2011 periods.

The best result, relative to the 2007–2011 period with Rotation Forest, has been obtained with the set of features shown in Table 5. The best set of parameters for Rotation Forest technique was Random Forest as the tree based method with 50 decision trees, 5 features for random selection and 7 as the maximum depth.

Table 6 compares the results of 15 trending terms obtained from both models. These terms were selected based on the time series trend analysis. In this table, the real TF-IDF of each term is compared with the predicted value from the time series prediction model and the results of the proposed approach. The prediction technique was Rotation Forest for the period 2007–2011 (the best prediction results presented, as shown in Table 4).

The accuracy gain displayed in Table 6 is a sample of the trend analysis improvement when including social network features. The experimental results show that the error produced by the proposed approach corresponds, in average, to only 17% of the error produced by the time series regression model and 18% of the error produced by the time series supervised learning methods which do not consider social network features.

In order to verify the quality of the proposed approach to identify trends over longer periods, additional experiments have been conducted fixing the dataset training period between years 1991 and 2005 and varying the prediction periods between the years 2006 and 2011 for testing. Only SVM and Rotation Forest have been employed in these experiments, as they yielded the best results in the previous experiments. Table 7 shows the results. As expected, the error rates increase with time. However,

Table 6 Results comparison for the 15 first trends of the time series prediction model in 2012

Term	Real	Time series regression	Error	Time series supervised learning	Error	Proposed	Error
Service discovery	135.17	441.52	306.35	76.74	58.43	123.39	11.77
Based approach	155.19	424.16	268.97	94.21	249.40	161.10	5.91
Information systems	147.32	334.29	186.97	34.76	182.08	148.37	1.05
Supply chain	174.31	298.37	124.06	28.60	145.71	143.96	30.35
Web services	225.28	297.74	72.46	35.14	190.14	201.05	24.23
Product line	174.99	291.57	116.57	306.69	481.68	154.73	20.26
Motion estimation	107.78	274.36	166.58	66.95	174.73	99.00	8.78
Social network	249.05	269.42	20.38	78.65	327.70	198.94	50.11
Business process	131.75	240.09	108.34	132.50	264.25	119.61	12.14
Time series	150.79	217.76	66.97	45.29	196.08	147.03	3.76
Neural network	213.36	178.86	34.51	352.45	565.81	198.85	14.51
Sign language	108.21	176.83	68.62	31.24	76.97	101.69	6.52
São Paulo	191.93	172.84	19.09	120.42	71.51	145.79	46.15
Genetic programming	128.25	156.64	28.39	24.07	104.18	107.98	20.26
Routing problem	101.11	147.16	46.05	94.50	195.61	83.75	17.36

Table 7 RAE for tests of short, medium, and long terms for the models trained in the period between 1991 and 2005

Method	2006	2007	2008	2009	2010	2011
RF	60.65%	62.25%	64.21%	65.54%	64.31%	67.87%
SVM	55.91%	57.43%	57.77%	58.70%	57.28%	59.29%

the errors do not increase dramatically for longer periods. Comparing these results with those obtained from time series regression methods presented in Table 2 one observes that the error rates are still lower.

Conclusion

Approaches that consider only the historical behavior of the analyzed object have been widely employed for trend prediction. However, the contents generated by people are clearly influenced by their connections. How information spreads is an important factor that can be considered in prediction. Intending to fill this gap, we presented a new approach for trend analysis incorporating the social network information to a content-based trend analysis model. The proposed approach achieved better results than the standard time series-based models. In addition to simple prediction techniques, such as linear regression, we applied more robust techniques that resulted in even more accurate models. As we supposed, these findings cast light on the issue of trend prediction. Information content and the characteristics of their social structure can be combined to improve the explanation of the information temporal behavior.

This work explored a concept still little studied and, thus, some shortcomings remain to be addressed. The dynamics of the social network is one of them. We worked with a fixed time window to the social network modeling. However, slicing the time interval probably would improve the prediction models by capturing the transient characteristics over time in the social structures. Another improvement could be achieved by grouping the extracted terms by topics, which can be more relevant than analyzing each term alone.

In conclusion, we found out that looking at the social structure of data sources alongside the main analyzed data can help better understanding the information temporal behavior.

Endnotes

[1] http://lattes.cnpq.br/.

[2] https://www.r-project.org/.

Acknowledgements
This work was partially funded by FAPESP, CAPES, and CNPq.

Authors' contributions
CCT developed, tested and validated the approach presented in this paper. LAD was Caio's advisor and contributed in the specification of the approach and the design of the experiments. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 26 July 2016 Accepted: 19 May 2017
Published online: 07 June 2017

References

1. Abe H, Tsumoto S (2009) Evaluating a method to detect temporal trends of phrases in research documents. In: 2009 8th IEEE International Conference on Cognitive Informatics. IEEE. pp 378–383. doi:10.1109/ICSMC.2009.5345958
2. Altshuler Y, Pan W, Pentland AS (2012) Trends prediction using social diffusion models. In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer, Berlin Heidelberg. pp 97–104. doi:10.1007/978-3-642-29047-3_12
3. Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. In: Proceedings of the 21st international conference on World Wide Web. ACM. pp 519–528. doi:10.1145/2187836.2187907
4. Cimenler O, Reeves Ka, Skvoretz J (2014) A regression analysis of researchers social network metrics on their citation performance in a college of engineering. J Informetrics 8(3):667–682. doi:10.1016/j.joi.2014.06.004
5. Clauset A, Newman ME, Moore C (2004) Finding community structure in very large networks. Physical review E 70(6):066 111
6. Digiampietri LA, Alves CM, Trucolo CC, Oliveira RA (2014) Análise da rede dos doutores que atuam em computação no Brasil. In: CSBC 2014 - BRASNAM. pp 33–44
7. Digiampietri LA, Mena-chalco JP, Melo POV, Malheiros AP, Meira DNO, Franco LF, Oliveira LB (2014) BraX-Ray: an x-ray of the Brazilian computer science graduate programs. Plos-ONE 9(4):e94541
8. Glanzel W, Schubert A (2004) Analysing scientific networks through coauthorship. In: Handbook of quantitative science and technology research. Kluwer Academic Publishers. pp 257–276. doi:10.1.1.86.4083
9. Hamilton JD (1994) Time series analysis, vol 2. Princeton university press, Princeton. ISBN: 9780691042893
10. Lemieux V, Ouimet M (2008) Análise Estrutural das Redes Sociais. Instituto Piaget
11. Moed HF, Glänzel W, Schmoch U (2004) Editors' introduction. In: Handbook of quantitative science and technology research. Springer Netherlands. pp 1–15
12. Nakagawa H, Mori T (2002) A Simple but Powerful Automatic Term Extraction Method. In: COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology - Volume 14, COMPUTERM '02. Association for Computational Linguistics, Stroudsburg. pp 1–7. doi:10.3115/1118771.1118778
13. Pan W, Aharony N, Pentland A (2011) Composite social network for predicting mobile apps installation. In: AAAI. arXiv:1106.0359
14. Pandit S, Yang Y, Chawla NV (2012) Maximizing information spread through influence structures in social networks. In: 2012 IEEE 12th International Conference on Data Mining Workshops. IEEE. pp 258–265. doi:10.1109/ICDMW.2012.140
15. Poblacion D, Mugnaini R, Ramos L (2009) Redes sociais e colaborativas em informação científica, 1st ed. Angellara Editoras,Sao Paulo
16. Pourkazemi M, Keyvanpour M (2013) A survey on community detection methods based on the nature of social networks. Iccke 2013 5(1):114–120. doi:10.1109/ICCKE.2013.6682855
17. Prell C (2012) Social network analysis history, theory & methodology, Los Angeles London SAGE
18. Scott J (2009) Social network analysis: a handbook, 2nd ed. SAGE. doi:10.4135/9781446294413
19. Singh S, Mishra N, Sharma S (2013) Survey of various techniques for determining influential users in social networks. In: Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), 2013 International Conference on. pp 398–403. doi:10.1109/ICE-CCN.2013.6528531
20. Teixeira LA, de Oliveira ALI (2009) Predicting stock trends through technical analysis and nearest neighbor classification. In: 2009 IEEE

International Conference on Systems, Man and Cybernetics. IEEE. pp 3094–3099. doi:10.1109/ICSMC.2009.5345944
21. Trucolo CC, Digiampietri LA (2014) Trend Analysis of the Brazilian Scientific Production in Computer Science. FSMA 14:2–9
22. Trucolo CC, Digiampietri LA (2014) Uma Revisão Sistemá. In: X Simpósio Brasileiro de Sistemas de Informação (SBSI 2014). Londrina. pp 639–650
23. Wang D, Wen Z, Tong H, Lin CY, Song C, Barabási AL (2011) Information spreading in context. In: Proceedings of the 20th International Conference on World Wide Web, WWW '11. ACM, New York. pp 735–744. doi:10.1145/1963405.1963508. [http://doi.acm.org/10.1145/1963405.1963508]
24. Wasserman S, Faust K (2009) Social network analysis: methods and applications. 19th ed. Social network analysis: methods and applications
25. Wasserman S, Galaskiewicz J (1994) Advances in social network analysis research in the social and behavioral sciences. SAGE. doi:10.4135/9781452243528