Research Online Publishing
www.ronpub.com

Open Access

Open Journal of Web Technologies (OJWT)
Volume 1, Issue 1, 2014
http://www.ronpub.com/journals/ojwt
ISSN 2199-188X

SIWeb: understanding the Interests of the Society through Web data Analysis

Marco Furini [A], Simone Montangero [B]

[A] Dipartimento di Comunicazione ed Economia, Università di Modena e Reggio Emilia, Viale Allegri 9, 42121 Reggio Emilia, Italy, marco.furini@unimore.it
[B] Institut für Quanteninformationsverarbeitung, Universität Ulm, D-89069 Ulm, Germany, simone.montangero@uni-ulm.de

Abstract

The high availability of user-generated contents in the Web scenario represents a tremendous asset for understanding various social phenomena. Methods and commercial products that exploit the widespread use of the Web as a way of conveying personal opinions have been proposed, but a critical thinking is that these approaches may produce a partial, or distorted, understanding of the society, because most of them focus on definite scenarios, use specific platforms, base their analysis on the sole magnitude of data, or treat the different Web resources with the same importance. In this paper, we present SIWeb (Social Interests through Web Analysis), a novel mechanism designed to measure the interest the society has on a topic (e.g., a real world phenomenon, an event, a person, a thing). SIWeb is general purpose (it can be applied to any decision making process), cross platforms (it uses the entire Webspace, from social media to websites, from tags to reviews), and time effective (it measures the time correlation between the Web resources). It uses fractal analysis to detect the temporal relations behind all the Web resources (e.g., Web pages, RSS, newsgroups, etc.) that talk about a topic and combines this number with the temporal relations to give an insight of the the interest the society has about a topic. The evaluation of the proposal shows that SIWeb might be helpful in decision making processes as it reflects the interests the society has on a specific topic.

Type of Paper and Keywords

Regular research paper: *Webspace analysis, social impact, society-web relation, trend analysis, fractal analysis.*

1 Introduction

The Internet and mobile technologies have been the primary force behind the ecosystem composed of blogs, microblogs, forums, wikis and social networks (just to name a few), where users, consumers, voters, business, governments and organizations produce more and more contents [10, 12, 31, 40].

The high rate at which people produce contents makes this scenario an important source of information to look at when analyzing the pulse of the society, as user-generated contents represent a tremendous asset for understanding various social phenomena, from extremism to social activism and from consumer sentiment to marketing intelligence [6, 9].

Traditionally, the task of understanding the interests of the society was accomplished through opinion polls, but this methodology is costly and time consuming to conduct. Therefore, researchers are proposing methods that exploit the widespread use of the Web as a way of conveying personal opinions with the aim of understanding the pulse of the society through analysis of the Web data. The rationale behind this approach is that the Web is a modern version of the ancient Greek Agora, where peo-

ple gathered together to do commercial and administrative activities, to discuss politics and philosophy, to participate to social and religious events, to understand and influence society. Indeed, Internet and mobile technologies are making the Web a significant representation of our society, as they create new spaces of freedom, allow users to express their opinions, facilitate interpersonal relationships, encourage the creation of collaborating collectivities and modify the way to conduct business.

Society and Web are so strongly linked that they affect each other: when something happens in the real world it is very likely that few seconds later someone writes about it in the Webspace. For instance, people use social media during and in response to anticipated and unanticipated events like natural disaster, disease outbreaks, speeches, elections and crises. Web users are akin to physical sensors, creating a global network of measurement capabilities: when something happens, users receives stimulus and they communicate through the system, other people receive the message and communicate through the system and the process goes on [8]. In this way, people create a network and contribute to all kinds of dynamic dialogs by sharing their expertise and opinions [40].

The large availability of user-generated contents provides a wealth of opportunities for understanding users' preferences, assessments, and opinions about contents, services, brands, people, events, etc. For instance, politicians may gauge public opinion on policies and/or political positions; government could get early clues about disease spreading and could plan appropriate countermeasure, corporations may find influential blogger to promote their products. In few words, the Web is becoming an essential component of the next-generation business intelligence platform.

In the literature, different proposals analyze the Web to get insights of what happens in the society: some focus on the blogosphere, other on specific platforms like Twitter and Facebook, and the applications vary from investigating general people' concerns/opinions to measure Hollywood stars' notoriety, from understanding politicians' popularity to identify consumers' opinions, (e.g., [2, 3, 5, 7, 8, 11, 13, 20, 23, 26, 35, 37]). The main limitations to most of these approaches are: i) they focus on definite scenarios like marketing and therefore an effective general-purpose approach is missing, ii) they analyze specific platforms like the Blogosphere, Twitter or Facebook, and therefore results represent only a portion of the society, iii) their analysis is mainly based on the sole magnitude of data and therefore it is easy to maliciously affect the input data to produce biased results, iv) they give the same importance to all the Web resources and therefore very old unrelated Web resources are considered similar to very recent and correlated Web resources. Needless to say, this may produce biased results.

Motivated by the limitations of current approaches, in this paper we propose SIWeb (Social Interests through Web Analysis), a novel mechanism to understand the interests of the society toward a topic (e.g., a real world phenomenon, an event, a person, a thing, an idea, etc.). The mechanism is designed with the following constraints: **General purpose**: it should not focus on specific scenarios, but it must be able to be used in any decision making process (e.g., marketing, society opinions, trend discovery, etc.); ii) **Cross platforms**: it should not use data of specific platforms, but it should use data of the entire Web space from social media to websites, from tags to reviews; iii) **Time effective**: it should not use the sole magnitude of data, but it has to detect and measure the temporal relations among the Web resources that talk about a specific topic.

The idea behind SIWeb is to use fractal analysis to detect and measure the temporal correlations among all the Web resources that talk about a particular topic and to combine the temporal correlations with the absolute number of Web resources that talk about the topic. The output of SIWeb is an index that gives an insight of the interests the society has about a specific topic. To evaluate our proposal, we considered different scenarios (politics, sports and cars) and we compare SIWeb against other methods. The results show that SIWeb better reflects the interests of the society and therefore it might be an helpful mechanism to measure the present and future interests around any topic.

The remainder of this paper is organized as follows. In Section 2 we overview works in the area of Web and Society; Section 3 presents details of the SIWeb proposal, whereas its evaluation is shown in Section 4. Conclusions are drawn in Section 5.

2 Related Works

Recently, in the literature many proposals focused on using Web data for different purposes. In the following, we present some of these proposals grouped in four different categories: prediction, people opinions, marketing and geosocial events.

- **Prediction**: Asur et al. [4] use the Twitter messages to forecast box-office revenues for movies. In particular, they constructed a linear regression model for predicting box-office revenues of movies in advance of their release. The obtained results showed that there is a strong correlation between the amount of attention a given topic has and its ranking in the future. Chi et al. [7] analyze the Blogosphere and propose a trend analysis technique based on the

singular value decomposition; Goel et al. [17] use the search query volume to predict consumer activities, such as attending movies and purchasing music or video games. The obtained results showed that in films, video games and music the search counts are highly predictive of future outcomes. Gruhl et al. [20] use the volume of blogs or link structures to predict the trend of product sales. Liu et al. [22] study the predictive power of opinions and sentiments expressed in blogs, in order to predict product sales performance; Glance et al. [26] propose a mechanism to discover trends inside the Blogosphere by using data mining techniques.

- **People opinions** Fukuhara et al. [11] describe a system that counts the number of blog articles containing a specific word so as to understand concerns of people; Merhav et al. [24] analyze the Blogosphere with natural language processing tools in order to provide a better understanding of the society. Indeed, their proposal extracts the relationships among entities, facts, ideas, and opinions. Ni et al. [27] propose a machine learning method for classifying informative and affective articles inside the Blogosphere; Weng and Lee [38] apply wavelet analysis to Twitter messages in order to detect bursts of word usage. The method is useful to analyze large events like sports or elections.

- **Marketing** Agarwal et al. [1] propose a method to identify influential bloggers into the blogosphere. Agrawal et al. [2] and Gamon et al. [14] have also conducted research in opinion mining for marketing purposes in the domains of newsgroup and blogs. Morinaga et al. [25] present an approach that automatically mines consumer opinions about target products from Web pages, in order to obtain the reputation of the products.

- **Geosocial events** Lee and Sumiya [21] as well as Pozdnoukhov and Kaiser [29] present methods to detect unusual geosocial events by measuring the spatial and temporal regularity of Twitter streams. Sakaki et al. [33] propose a natural disaster alert system using Twitter users as virtual sensors.

Also commercial blog sites and Web search engines are offering services that aim at understanding society through Web data analysis. The Webfountain project [19] uses Web mining techniques for market intelligence and is based on massive server clusters; Google Trends[1] analyzes a percentage of Google web searches to determine how many searches have been done for a specific term compared to the total number of Google searches done during that time [18]; different studies use Google Trends to predict future events (e.g., influenza cases [15] or stock market [30]).

Motivated by the fact that the approaches designed to investigate people' opinions are usually based on specific platforms (e.g., Blogosphere, Twitter, Web searches), analyze the Web data with a magnitude analysis and don't consider the temporal correlations in the Web data, in this paper we propose a novel mechanism that aims at understanding the interests of the society through Web data analysis.

Our approach is different: SIWeb is designed to be a general purpose approach to understand the interests of the society through Web data analysis, it is based on the whole Webspace and not on a subset of it (like the Blogosphere, Web searches or specific social media) and, most important, it uses fractal analysis to detect and measure the temporal relations among the Web resources related to a specific topic in order to differentiate their importance. Indeed, it is important to consider temporal correlations among Web resources, as the Web is a time evolving scenario where the number of Web resources that talk about a specific topic is different from time to time; the more these Web resources are temporarily correlated, the more the topic reflects an interest of society. In particular, correlations that survive long enough are likely to create a network of Web resources. If this happens, the network will likely respond to subsequent stimulus (new events related to the topic in a similar "correlated" way). Conversely, if the network is not sufficiently correlated, it will eventually vanish and disappear, and the response to subsequent stimulus will be negligible. Although fractal analysis has been extensively employed in diverse scientific, sociological, and philosophical areas of research, and is used to describe physical, visual, acoustic, and chemical processes, and biological, weather, and financial systems [39], to the best of our knowledge, SIWeb is the first mechanism that uses fractal analysis to understand the interests of the society through Web data analysis.

3 THE SIWEB PROPOSAL

The Web is becoming the principal provider of news and opinions, and the society-Web relation can be exploited to discover people concerns, opinions, and trends by analyzing the Webspace. In this section, we present details of SIWeb (Society Interests through Web Analysis), a novel mechanism designed to understand the interests of the society through Web data analysis. The motivation behind our proposal is that current approaches might provide a misleading perception of what's happening in society as they: i) focus on specific platforms, (e.g., Twit-

[1] http://www.google.com/trends

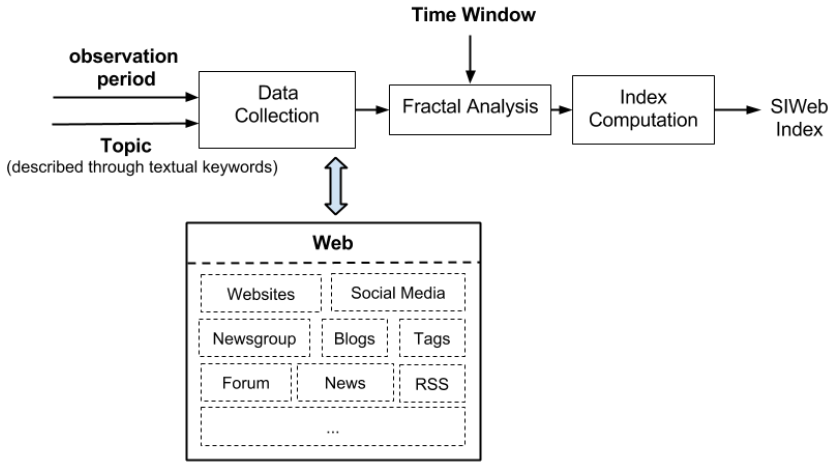

Figure 1: SIWeb Architecture. The entire Webspace and fractal analysis are used to understand the interests the society has on a specific topic.

ter, Facebook, Blogosphere, etc.) and therefore they may produce biased results as they analyze data generated by a portion of the society; ii) use the sole magnitude of data and therefore it is easy to maliciously alter the input data; iii) do not consider the temporal correlations in the Web data and therefore they may produce results that do not reflect the interests of the society.

Before presenting the details of our proposal, we recall here that, in the following, the term "Web resource" refers to any resource (e.g., a Web page, an RSS feed, a discussion in a newsgroup, etc) that can be accessible, and retrievable, in the Web, and that correspond to a specific topic; the term "topic" refers to someone or something that people talk about (e.g., a natural event, a person, a thing, an idea, etc.) in the society and that is possible to describe through a set of textual keywords.

SIWeb is designed with the following constraints: i) **General purpose**: it should not focus on specific scenarios, but it must be able to be used in any decision making process (e.g., marketing, society opinions, trend discovery, etc.); ii) **Cross platforms**: it should not use data of specific platforms, but it has to use data of the entire Web space from social media to websites, from tags to reviews; iii) **Time effective**: it should not use the sole magnitude of data, but it has to detect and measure the temporal relations among the Web resources that talk about a specific topic.

Figure 1 presents the SIWeb architecture: the "data collection" module accesses to the Web and harvests the number of Web resources that talk about a topic (identified by a set of textual keywords) in a specific period of time (observable period); the module produces a time-series of values that represent the number of Web resources talking about a particular topic as a function of time; Fractal analysis is then applied to the time-series and the results are combined with the absolute number of Web resources that talk about the considered topic in order to compute the SIWebIndex. Indeed, fractal analysis, through the computation of the fractal dimension, gives an insight of the amount of correlations present in the network of Web resources. Technically speaking, it gives a fast insight of the "system" that generated the time-series (i.e., it tells whether the system is regular, random, or something in between). An example of regular system is represented by a single blogger who posts different messages about the same topic. Although the number of Web resources talking about the topic smoothly increases, this increasing number does not reflect a growing of interest in topic by society; it simply represents a blogger very fond of the topic. An example of a random system is represented by several Web resources without any correlations that talk about the same topic (e.g., people who post messages about the same topic but do not relate each other). It is worth noting that, in many scenarios, the knowledge of the system brings considerable benefits: for instance, it may be useful to help predicting the near future behavior of earthquakes, or the stock market trend [34, 36].

Anything in between a regular and a random system means that the network of Web resources that generated the sequence is correlated and thus it will likely cause other people to become part of the network. As a result, more people are interested in the topic described by the set of keywords. To better clarify, let us consider a simple example: the success of a TV-series. An ensemble of fans can be triggered by the pilot episode so as to

form a group of people interested in the TV-series. In this case, the group of fans is a correlated network as they talk almost every day about the TV-series. In fact, when a new episode is aired this group of fans will easily be the first to talk about it, and it is likely that they will cause other people to become fans; this means that the network grows, as additional people become part of the network. If the network is correlated enough, for years to come there will be people speaking, reading, listening, and talking about the TV-series.

Note that, by periodically using SIWeb, it is possible to obtain a SIWebIndex time-series that shows how the interest that society has on a specific topic changes with time.

3.1 Data collection

This module is in charge of retrieving the number of Web resources (e.g., Web pages, social media, News feeds, blogs, tags) that talk about a topic on a specific period of time. It takes in input two parameters: the set of textual keywords that describe the topic and the observable period. By measuring the number of Web resources that talk about the topic in several consecutive days, the module produces a time-series that describes the number of Web resources talking about the topic over time. In essence,

$$\Gamma_{Web}(S) = \{x_1, \ldots, x_N\} \quad (1)$$

represents the time-series of N consecutive and periodic measures of the number of Web resources found when searching for the set S of textual keywords, and x_i is the number of Web resources talking about the topic at time i.

3.2 Fractal analysis

We mentioned that, given a time-series, fractal analysis gives a fast insight of the "system" that generated the sequence of values. This is achieved with the computation of the so-called fractal dimension, which allows discerning whether the system that generated the time-series is regular or random. Roughly, a regular system produces smooth changes in the sequence of values, whereas a random system produces highly irregular changes in the sequence. Note that, in our scenario, the system is composed of all the Web resources described through the set S of textual keywords, and the time-series is the one generated by the data collection module (i.e., $\Gamma_{Web}(S)$).

To compute the fractal dimension D of a sequence of N samples (e.g., the sequence of values collected in different time points), SIWeb uses the box counting algorithm [32]: a grid of square boxes of size L^2 (see Figure 2) covers the data. The number $M(L)$ of boxes needed to cover the curve is recorded as a function of the box size L. The (fractal) dimension D of the curve is then defined as

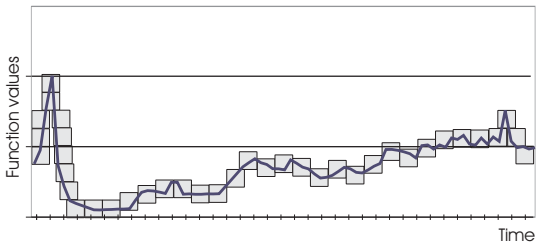

Figure 2: The Fractal Dimension of a curve is obtained by covering the curve with $M(L)$ squares of dimension L.

$$D = -\lim_{L \to 0} \log_L M(L). \quad (2)$$

One finds $D = 1$ for a straight line (regular system), whereas $D = 2$ for a random curve. Indeed, eventually a random curve covers uniformly the whole plane. Any other value of D in between of these integer values is a signal of the fractality of the curve. This algorithm can be modified using rectangular boxes of size $L \times \Delta_i$ (Δ_i is the largest excursion of the curve in the region L). Then, the number

$$M(L) = \frac{\sum_i \Delta_i}{L} \quad (3)$$

is computed; for any curve a region of box lengths $L_{min} < L < L_{max}$ exists where $M \propto L^{-D}$. Outside this region one either finds $D = 1$ or $D = 2$: The first equality ($D = 1$) holds for $L < L_{min}$ and is due to the coarse grain artificially introduced by any discrete time series. The second one ($D = 2$) is obtained for $L > L_{max}$ and is due to the finite length of the analyzed time series. The exponent D is then extracted from the function $M(L)$ by means of a fit in the region L_{min}, L_{max}. It is important to highlight that the fit result might depend on the choice of the boundaries L_{min}, L_{max}. For this reason, L_{min}, L_{max} are chosen by an adaptive algorithm that aims at minimizing the introduced error. As shown in the experimental results, the error, comprehensive of the errors introduced by the fitting procedure, is kept small enough to produce interesting SIWeb indexes.

The fractal dimension measures the degree of correlations in a time series, as shown for example in [16]. Given a real stochastic process $x(t)$ as a function of time with zero mean and unit variance, i.e. $E[x(t)] = 0$ and $E[x^2(t)] = 1$, the correlations present in the time series are defined as the expectation value as a function of the distance h, $C(x, h) = E[x(t)x(t + h)]$. In the case of a

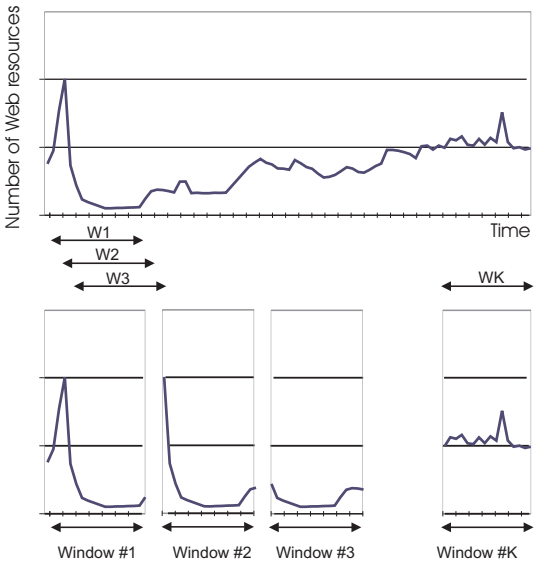

Figure 3: The time-series representing the number of Web resources talking about a specific topic is split into several overlapping and consecutive windows. Fractal analysis is then applied to each window to compute the fractal dimension.

stationary Gaussian random process, one can show that if the correlations in the time series are $C(h) = 1 - |h|^{\beta}$ as $h \to 0$ for some $\beta \in (0, 2]$, then the fractal dimension is related to the exponent β as follows:

$$D = 2 - \frac{\beta}{2}.$$

Thus, the faster the decaying of the correlation the lower the fractal dimension is: for example $D = 2$ corresponds to the case of very slow decaying correlations $\beta = 0$, while $D = 1$ to fast decaying correlations $\beta = 2$.

To better appreciate how this temporal correlation changes with time, SIWeb considers the N collected samples of the time-series $\Gamma_{Web}(S)$ as several overlapping and consecutive time windows of length z (e.g., in our experiments we consider z equal to 14 and 30 days). Figure 3 better shows the process. As a result, if $N > z$, N samples produces $N - z + 1$ time windows, with z the length of the time-window. Roughly, each time window represents a *snapshot* of the Web and its analysis allows discovering the temporal correlation of the Web resources in that particular time-window.

SIWeb computes the fractal dimension of each window, thus it produces a sequence of values $D_z^{\Gamma_{Web}(S)}$ composed of $(D_{(1,z)}(S), D_{(2,z)}(S), ..., D_{(N-z+1,z)}(S))$, where $D_{(i,z)}(S)$ is the fractal dimension of the $i - th$ time window of length z.

It is worth noting that the sequence $D_z^{\Gamma_{Web}(S)}$ is not sufficient to understand the interests the society has on the topic described by the set of keywords S. Indeed, the interest is also represented by the total number of Web resources talking about it. Therefore, SIWeb combines together $D_z^{\Gamma_{Web}(S)}$ and $\Gamma_{Web}(S)$, as shown in the following.

3.3 SIWebIndex Computation

The SIWebindex is a combination between the results obtained with the fractal analysis and the number of Web resources. Since SIWeb considers the N collected samples of the time-series as a sequence of $N - z + 1$ time windows, first it computes the average number of Web resources described by the set of keywords S in any time window and then it combines temporal correlation and number of Web resources to compute SIWebIndex.

In particular, by denoting the $i-th$ time window with T_i, the average number of Web resources described by the set of keywords S in the $i - th$ window of length z is computed as follows:

$$W_i^z(S) = \sum_{j \in T_i} x_j / |T_i| \tag{4}$$

The SIWebIndex of the $i-th$ window is a combination of $W_i^z(S)$ and of $D_{i,z}(S)$ and is computed as follows:

$$SIWebIndex_i^z(S) = \alpha \cdot \log W_i^z(S) + (1-\alpha) \cdot D_{i,z}(S) \tag{5}$$

where α is a parameter in the range [0..1] used to tune whether one wants to give more importance to the absolute number of Web resources or to the temporal correlations among these resources. Indeed, a low value of α gives more prominence to the absolute number of web resources, whereas a high value of alpha gives more importance to relationships among the Web resources.

By applying Equation (5) to all the $(N - z + 1)$ time windows, we find the sequence $SIWebIndex_1^z(S)$, $SIWebIndex_2^z(S)$,...,$SIWebIndex_{N-z+1}^z(S)$ that shows how the interests the society has on the topic described with the set of keywords S changed over time.

4 SIWeb Evaluation

To evaluate our proposal, we develop a prototype version of the SIWeb mechanism using the Google searching tools that allows retrieving Web results for a specific day. The research options have no restrictions with respect to country and language (any language and any country). However, it is worth highlighting that any other data collection tool can be used (e.g., another search engine, a developed one) and any restriction can be specified (e.g.,

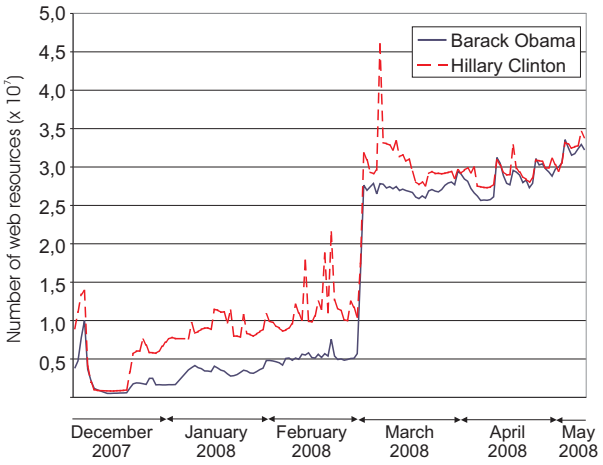

Figure 4: Number of Web resources talking about the Democratic candidates *Barack Obama* and *Hillary Clinton*.

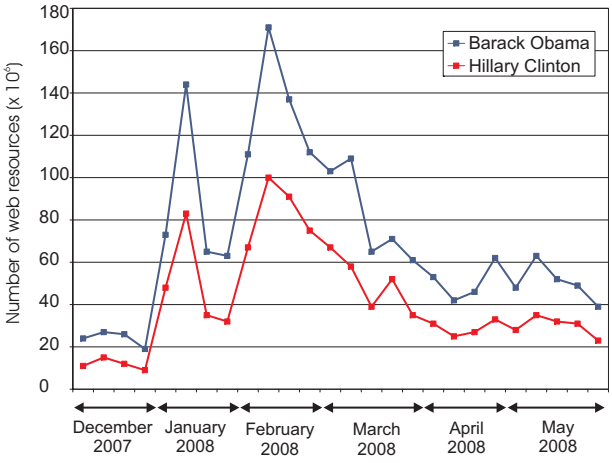

Figure 5: Number of Web searches related to the Democratic candidates *Barack Obama* and *Hillary Clinton*.

if someone is looking for the interests of the society in specific countries).

In the following, we present three different scenarios: politics (the 2008 US Democratic primaries elections and the 2008 US Presidential elections), sports (the interests around Tennis top players), and cars (the interests around some popular car manufacturers brands).

We mentioned that the SIWeb index is customizable by setting the value of z (the length of the time window where to compute the fractal dimension) and the value of α (the balance between the absolute number of web resources and the relations among the Web resources).

Far from proposing the best values of these parameters, in the following experiments, we consider two different values of z (14 and 30) and the value of 0.5 for the α parameter (same importance for the absolute number of Web resources and for the relation among these Web resources). Indeed, although we observed that these values produced reasonable results, it is worth noting that the tuning of these parameters is outside the scope of the paper and is left to the SIWeb users (as happen with any other data management and analysis tools).

4.1 Politics scenario

In 2008, the Democratic Party had to choose its nominee for President of the United States in the 2008. This is usually done using the so-called Presidential primaries process, a sequence of primary races and caucuses to be held in each of the fifty US states. According to several poll agencies, Hillary Clinton was the strongest candidate among the eight initial ones, and after the first races and caucuses (held in January 2008), it was clear, and unexpected, that the race was between Barack Obama and Hillary Clinton. The contest remained competitive for several months, and only at the beginning of June, Hillary Clinton withdraw and conceded the nomination to Barack Obama.

In the following, we investigate the scenario by looking at the number of Web resources and at the number of Web searches that were done every day during that period, and then we apply our proposal to this scenario.

Figure 4 reports the number of Web resources talking about the Democratic candidates *Barack Obama* and *Hillary Clinton*. At first glance, Hillary Clinton had a much larger number of resources that talked about her (only at the beginning of April, the two candidates had a comparable number of Web resources). It can also be noted the presence of peaks, which happen during primary election races or caususes. Also, it is interesting to note that the number of Web resources increased a lot at the beginning of March. A reasonable explanation is that at the beginning of March, the candidate John McCain got the Republican nomination, and hence all the media attentions began focusing mainly on the Democratic party. If one thinks that the number of Web resources represents the interest of the society, by looking at this chart, Hillary Clinton should have won every primary election race and caucus, but we know she did not.

Figure 5 reports the volume of Web searches related to the Democratic candidates *Barack Obama* and *Hillary Clinton*. Results show that the term "Barack Obama" has been entered in Web search engines many more times than the term "Hillary Clinton". If one thinks that the number of Web searches represents the interest of the society, by looking at this chart, Barack Obama should

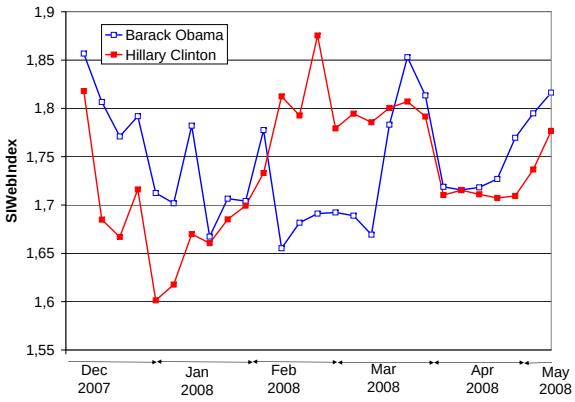

Figure 6: SIWebIndex related to the Democratic candidates "Barack Obama" and "Hillary Clinton", with a time window of 30 days and $\alpha = 0.5$. Higher values mean a higher interest in society. Errors are in the range [0-0.02].

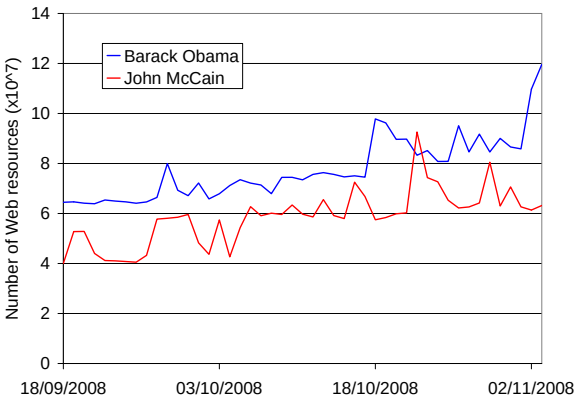

Figure 7: Number of Web resources related to the US Presidential candidates *Barack Obama* and *John McCain*.

have won every primary election race and caucus, but we know he did not.

These examples show that an analysis based on the simple magnitude of results (e.g., Web resources, Web engine searches) may represent a distorted reality, and therefore is not sufficient to understand the interests the society has on a specific topic.

Figure 6 reports the SIWebIndex computed with $\alpha = 0.5$ (same weight to correlations and to the magnitude of the web resources) and with $z = 30$ days. The experimental results have errors in the order $\epsilon = 0.02$. It is worth reminding that the period up to the end of January saw Barack Obama winning primary election contests (Iowa and S. Carolina) and getting interesting results in others (New Hampshire, Nevada, and Florida). The majority of the media defined these results as *unexpected*, but looking at the SIWebIndex, these results were not unexpected at all: in this period the SIWebIndex related to Barack Obama has been always higher than the one of Hillary Clinton. A second interesting period to analyze is February. In that period, all the media mentioned a possible withdraw of Hillary Clinton from the Presidential race. From Figure 6 it can be noted that the SIWebIndex related to Hillary Clinton has been always higher than the one related to Barack Obama: this shows that the society was more interested to Hillary Clinton. In the second half of March (when no primary election contests were scheduled), the interests around the two candidates decreased. When the primary election contests begun again, the interests of both increased, with the one about Barack Obama higher than the one of Hillary Clinton (it is to note that at the beginning of June 2008, Hillary Clinton withdraws from the Presidential race, and Barack Obama became the Democratic nominee for President of the United States).

This case-study shows that, while approaches based on the simple magnitude of results were not sufficient to understand society, the SIWebIndex better reflects the interests the society has on a specific topic.

Another interesting example is related to the 2008 US Presidential election that was held on November 4, 2008: people had to choose between Democratic Party nominee Senator Barack Obama and Republican Party nominee Senator John McCain. Barack Obama became the 56th US President.

Figure 7 reports the number of Web resources talking about *Barack Obama* and *John McCain* from mid September to November 3 (the period when the Presidential campaign became interesting). During this period, there were more Web resources talking about Barack Obama than about John McCain. It is interesting to observe that on October 16, John McCain almost reached Barack Obama, and even passed him on October 22. Looking at what happened in society, we observe that on October 16, John McCain was a guest of the "David Letterman Show" and the video became very popular on video sharing sites like YouTube; on October 22, media talked a lot about rumors related to expenses campaign of the John McCain's Vice-President. Also to note the impact that a speech held in St. Louis on October 18 had on Barack Obama. Note also how, as of November 3, the difference among the two candidates is quite clear.

Figure 8 reports the number of Web searches talking about *Barack Obama* and *John McCain* from mid September to November 3. Looking at the number of

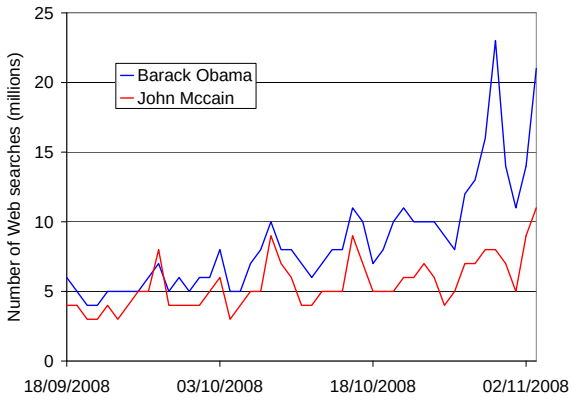

Figure 8: Number of Web searches related to the US Presidential candidates *Barack Obama* and *John McCain*.

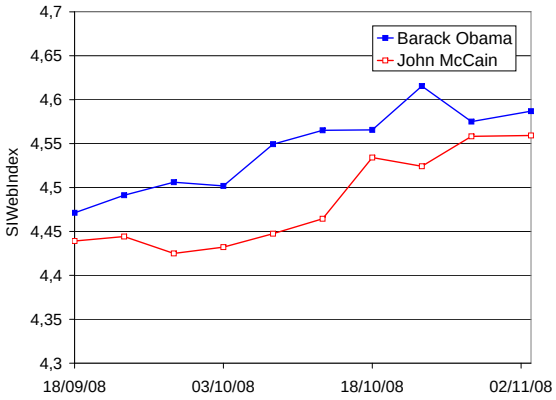

Figure 9: SIWebIndex related to "Barack Obama" and "John McCain", with a time window of 30 days and $\alpha = 0.5$. Higher values mean a higher interest in society. Errors are in the range [0-0.04].

Web searches, the term "Barack Obama" has been entered many more times than the term "John McCain".

In summary, approaches based on the magnitude of results clearly showed that Barack Obama was taking the lead over John McCain. However, according to several pool agencies, John McCain narrowed the gap from 15 to 7 points in the last week of the Presidential campaign [28]. This aspect is invisible in the approaches based on the magnitude of results.

Figure 9 reports the SIWebIndex computed with equation (5) with $\alpha = 0.5$ (same weight to correlations and to the magnitude of the web resources) and $z = 30$ days. The experimental results have errors in the order $\epsilon = 0.04$. It can be observed that the interests of the society were more focused on "Barack Obama". However, it is interesting to observe the last two weeks of the campaign: "John McCain" highly reduced the gap, which is exactly what poll agencies reported.

Once again, this scenario shows that the SIWebIndex better reflects the interests the society has on a specific topic.

4.2 The auto scenario

The automotive industry is one of the largest of all industries. With no doubts, the economic and financial crisis affected the automotive Industry. In such a scenario, it is interesting to observe what are the interests of the society toward some of the most popular cars manufacturers. To this aim, Figure 10 presents the absolute number of Web resources that talk about some of the most popular cars manufacturers (FIAT, Opel, Volvo, Chrysler, Toyota and Renault). Looking at the Figure, it is difficult to have a clear idea of the scenario: there are several peaks that usually correspond to specific events (e.g., the peak of the FIAT curve at the end of December corresponds to news about FIAT taking control of Chrysler, the peak of the Renault curve at the beginning of January corresponds to the presentation of a racecar model at the CES in Las Vegas). Figure 11 presents the SIWeb Index related to the considered cars manufacturers. The index is computed with $\alpha = 0.5$ and $z = 14$ days and the experimental results have errors in the order $\epsilon = 0.14$. We recall here that higher values mean a higher interest in society. It is interesting to observe the FIAT topic: the interest around the brand increased around the mid of December, where the first rumors about Chrysler acquisition began to appear on the news. This interest is transparent to the absolute number of Web resources: indeed, by looking at the sole number of Web resources one may infer that the peak at the end of December likely corresponds to an interest, but it is difficult to understand the interest in the days pre and post peak. Another interesting case is what happen to the Opel topic: on December 27, Reuters reported that *"General Motors Company has declared that its European unit Opel remains positive that sales growth in 2014 will be sufficient so that the company won't need to make further cost cuts"*. Although this news did not cause a considerable increment in the absolute number of Web resources that talk about the Opel topic, SIWeb measured a high correlation, showing that this news created an interest in the society.

4.3 The Tennis scenario

Tennis is played and watched by millions of people all over the world and top players are sponsored by top com-

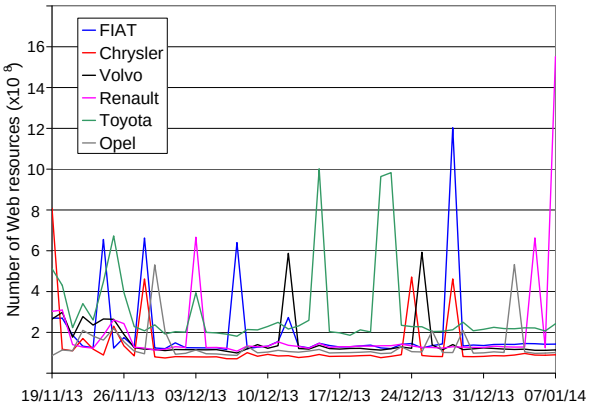

Figure 10: The absolute number of Web resources that talk about cars manufacturers like FIAT, Opel, Volvo, Chrysler, Toyota and Renault.

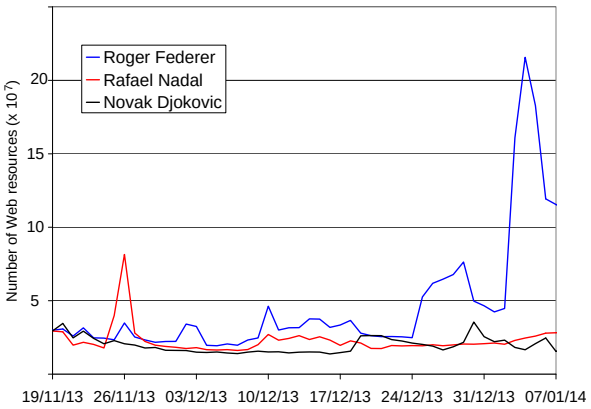

Figure 12: The absolute number of Web resources that talk about top tennis players like Roger Federer, Rafael Nadal and Novak Djiokovic.

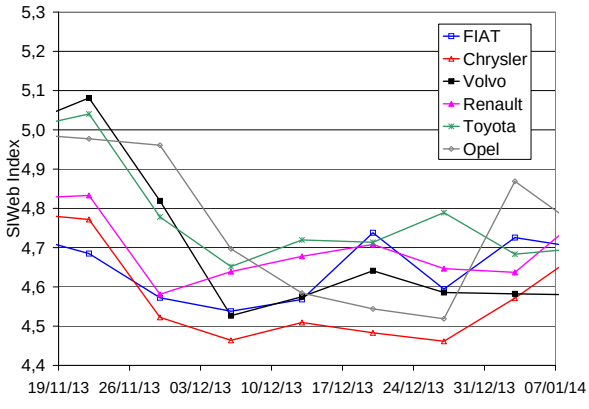

Figure 11: SIWeb index related to cars manufacturers like FIAT, Opel, Volvo, Chrysler, Toyota and Renault. Time window of 14 days and $\alpha = 0.5$. Higher values mean a higher interest in society. Errors are in the range [0-0.14].

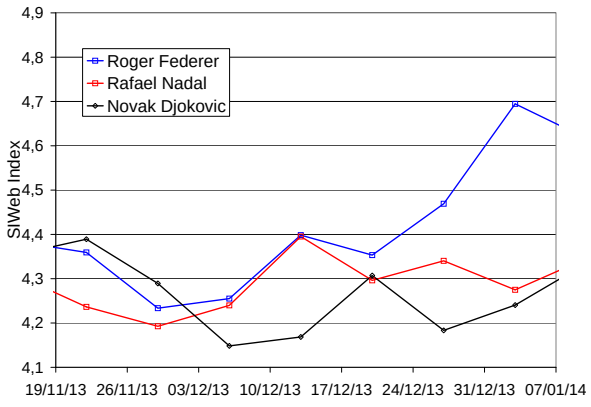

Figure 13: SIWeb index related to top tennis players. Time window of 14 days and $\alpha = 0.5$. Higher values mean a higher interest in society. Errors are in the range [0-0.11].

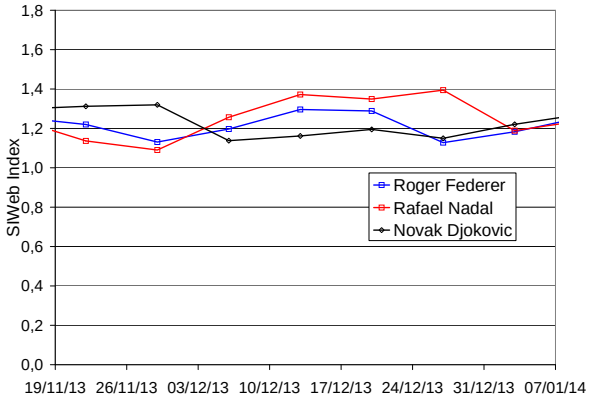

Figure 14: SIWeb index related to top tennis players. Time window of 14 days and $\alpha = 0$. Higher values mean a higher interest in society. Errors are in the range [0-0.11].

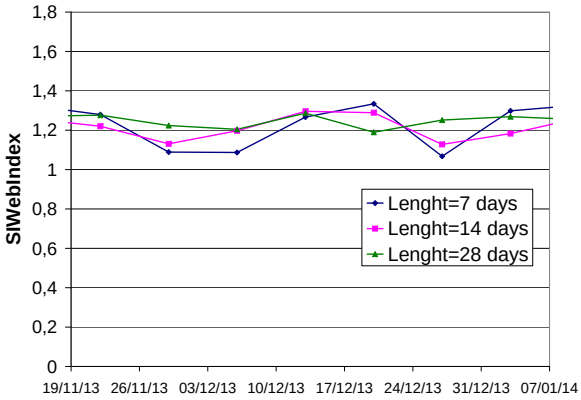

Figure 15: SIWebIndex ($\alpha = 0$) related to 'Roger Federer': the influence of the time window length. Errors are in the range [0-0.11].

panies. In this scenario, it is interesting to measure the interests around top player. This information will be very useful to top companies when deciding what player to sponsor.

Figure 12 presents the absolute number of Web resources that talk about top tennis players like Roger Federer, Rafael Nadal and Novak Djiokovic. Looking at the Figure, it seems that people are interested in Roger Federer (with the exception of the Rafael Nadal peak around the end of November, when the Spanish tennis player got the best Spanish athlete award). However, if we look at Figure 11 (the index is computed with $\alpha = 0.5$ and $z = 14$ days and the experimental errors are in the order of 0.11), it is interesting to observe that, in the first ten days, people were more interested in Novak Djokovic. This interest is transparent to the absolute number of Web resources.

4.4 Summary of Results

The experiments showed that an analysis based on the simple magnitude of results (e.g., Web resources, Web engine searches) may represent a distorted reality, and therefore is not sufficient to understand the interests the society has on a specific topic. Conversely, SIWeb produces an index that better reflects the interests the society has on a specific topic.

For completeness, we recall here that SIWeb requires to specify two different parameters in order to compute the SIWeb index: the length of the time windows where to compute the fractal dimension and the weight of the correlations among the Web resources that talk about a specific topic. In our experiments, we considered two different lengths of time windows (14 and 30 days) and an equal importance between the number of Web resources that talk about a specific topic and their correlation (e.g., α=0.5). To show how the parameters affect the SIWebindex, Figure 14 shows the SIWebIndex related to top tennis players computed by considering a time window of 14 days and by giving importance only to relations among the Web resources that talk about the top tennis players, i.e., $\alpha = 0$. The obtained results show that the Web resources that talked about Nadal were more correlated that the ones that talked about Federer or Djokovic, showing that people "talked" more about Djokovic (first ten days) and Nadal than about Federer. Figure 15 shows the SIWebIndex related to "Roger Federer" computed by considering $\alpha = 0$ and by varying the time window length (7, 14 and 28 days). The results computed with shorter windows have more variations compared to the results obtained by using longer windows. In fact, long time window length produces smooth results (e.g., if we consider a single window with a time length equal to the one of the time series, there would be only one result, the one representing the average correlation present in the entire series).

5 Conclusions

In this paper, we presented SIWeb (Social Interests through Web Analysis), a novel mechanism designed to understand the interests the society has on a specific topic. SIWeb is proposed to overcome some limitations of other approaches (e.g., usage of specific platforms, analysis based on the sole magnitude of data and temporal correlations of Web data not considered) that may produce biased results, and is designed to be: i) general

purpose, ii) cross platforms, and iii) time effective. It combines the number of Web resources that talk about a specific topic with the amount of correlations among these Web resources and gives an insight of the interests the society has on a specific topic. The evaluation of the proposal shows that SIWeb might be helpful in decision making processes as it reflects the interests the society has on a specific topic.

REFERENCES

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ser. WSDM '08. New York, NY, USA: ACM, 2008, pp. 207–218. [Online]. Available: http://doi.acm.org/10.1145/1341531.1341559

[2] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu, "Mining newsgroups using networks arising from social behavior," in *WWW '03: Proceedings of the 12th international conference on World Wide Web*. New York, NY, USA: ACM, 2003, pp. 529–535.

[3] E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer, "Trend detection through temporal link analysis," *J. Am. Soc. Inf. Sci. Technol.*, vol. 55, no. 14, pp. 1270–1281, 2004.

[4] S. Asur and B. A. Huberman, "Predicting the future with social media," *CoRR*, vol. abs/1003.5699, 2010.

[5] E. Bothos, D. Apostolou, and G. Mentzas, "Using social media to predict future events with agent-based markets," *Intelligent Systems, IEEE*, vol. 25, no. 6, pp. 50–58, 2010.

[6] H. Chen and C. C. Yang, "Special issue on social media analytics: Understanding the pulse of the society," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 41, no. 5, pp. 826–827, 2011.

[7] Y. Chi, B. L. Tseng, and J. Tatemura, "Eigen-trend: trend analysis in the blogosphere based on singular value decompositions," in *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*. New York, NY, USA: ACM, 2006, pp. 68–77.

[8] C. Corley, C. Dowling, S. Rose, and T. McKenzie, "Social sensor analytics: Measuring phenomenology at scale," in *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, 2013, pp. 61–66.

[9] E. Diaz-Aviles, C. Orellana-Rodriguez, and W. Nejdl, "Taking the pulse of political emotions in latin america based on social web streams," in *Web Congress (LA-WEB), 2012 Eighth Latin American*, 2012, pp. 40–47.

[10] S. Ferretti, M. Furini, C. E. Palazzi, M. Roccetti, and P. Salomoni, "WWW recycling for a better world," *Commununication of the ACM*, vol. 53, no. 4, pp. 139–143, 2010.

[11] T. Fukuhara, T. Murayama, and T. Nishida, "Analyzing concerns of people using weblog articles and real world temporal data," in *WWW 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, May 2005.

[12] M. Furini, "Users behavior in location-aware services: Digital natives vs digital immigrants," *Advances in Human-Computer Interaction*, vol. 2014, Hindawi Press, 2014, doi:dx.doi.org/10.1155/2014/678165.

[13] M. d. R. G. Mishne, "A study of blog search," in *ECIR '06: Proceedings of the 28th European Conference on Information Retrieval*. Springer Press, April 2006, pp. 289–301.

[14] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: Mining customer opinions from free text," 2005, pp. 121–132. [Online]. Available: http://dx.doi.org/10.1007/11552253_12

[15] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012–1014, 2009, doi:10.1038/nature07634. [Online]. Available: http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html

[16] T. Gneiting and M. Schlather, "Stochastic models that separate fractal dimension and the hurst effect," *SIAM Review*, vol. 46, no. 2, pp. 269–282, 2004. [Online]. Available: http://epubs.siam.org/doi/abs/10.1137/S0036144501394387

[17] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, "Predicting consumer behavior with web search," *Proceedings of the National Academy of Sciences*, 2010. [Online]. Available: http://www.pnas.org/content/early/2010/09/20/1005962107.abstract

[18] Google, "Google trend support," *https://support.google.com/trends/?hl=en #topic=4365599*, 2014.

[19] D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, and J. Zien, "How to build

a webfountain: An architecture for very large-scale text analytics," *IBM Syst. J.*, vol. 43, no. 1, pp. 64–77, 2004.

[20] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *WWW '04: Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA: ACM, 2004, pp. 491–501.

[21] R. Lee and K. Sumiya, "Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection," in *Proceedings of the 2Nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, ser. LBSN '10. New York, NY, USA: ACM, 2010, pp. 1–10. [Online]. Available: http://doi.acm.org/10.1145/1867699.1867701

[22] Y. Liu, X. Huang, A. An, and X. Yu, "Arsa: a sentiment-aware model for predicting sales performance using blogs," in *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2007, pp. 607–614.

[23] D. Mahata and N. Agarwal, "What does everybody know? identifying event-specific sources from social media," in *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on*, 2012, pp. 63–68.

[24] Y. Merhav, F. Mesquita, D. Barbosa, W. G. Yee, and O. Frieder, "Extracting information networks from the blogosphere," *ACM Trans. Web*, vol. 6, no. 3, pp. 11:1–11:33, Oct. 2012. [Online]. Available: http://doi.acm.org/10.1145/2344416.2344418

[25] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the web," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2002, pp. 341–349.

[26] T. T. N. S. Glance, M. Hurst, "Blogpulse: Automated trend discovery for weblogs," in *WWW '04: Proceedings of the 13th international conference on World Wide Web*. New York, NY, USA: ACM, 2004.

[27] X. Ni, G.-R. Xue, X. Ling, Y. Yu, and Q. Yang, "Exploring in the weblog space by detecting informative and affective articles," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 281–290.

[28] PewResearch, "Mccain narrows gap," in *http://www.people-press.org/files/legacy-pdf/468.pdf*, November 2008.

[29] A. Pozdnoukhov and C. Kaiser, "Space-time dynamics of topics in streaming text," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, ser. LBSN '11. New York, NY, USA: ACM, 2011, pp. 1–8. [Online]. Available: http://doi.acm.org/10.1145/2063212.2063223

[30] T. Preis, D. Reith, and H. E. Stanley, "Complex dynamics of our economic life on different scales: insights from search engine query data," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. Vol.368, no. No.1933, pp. 5707–5719, 2010.

[31] M. Roccetti, S. Ferretti, C. E. Palazzi, M. Furini, and P. Salomoni, "Riding the web evolution: from egoism to altruism," in *Proceedings of the IEEE Consumer Communication & Networking 2008 (CCNC2008)*, January 2008, pp. 1123–1127.

[32] A. S. Sachrajda, R. Ketzmerick, C. Gould, Y. Feng, P. J. Kelly, A. Delage, and Z. Wasilewski, "Fractal conductance fluctuations in a soft wall stadium and a sinai billiard," *Phys. Rev. Lett.*, vol. 80, p. 1948, 1998. [Online]. Available: http://www.nld.ds.mpg.de/downloads/publications/p1948_1.pdf

[33] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 851–860. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772777

[34] S. Selvaratnam and M. Kirley, "Predicting stock market time series using evolutionary artificial neural networks with hurst exponent input windows," in *Proceedings of AI 2006: Advances in Artificial Intelligence*. Springer Press, December, 4–8 2006, pp. 617–626.

[35] V. Singh, P. Waila, R. Sadat, R. Piryani, and A. Uddin, "Computational analysis of thematic blog data for sociological inference mining," in *Applied Computational Intelligence and Informatics (SACI), 2013 IEEE 8th International Symposium on*, 2013, pp. 293–298.

[36] D. Sornette, in *Why Stock Markets Crash: Critical Events in Complex Financial Systems*. Princeton University Press, 2008.

[37] P. Waila, V. Singh, and M. Singh, "Blog text analysis using topic modeling, named entity recog-

nition and sentiment classifier combine," in *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, 2013, pp. 1166–1171.

[38] J. Weng and B.-S. Lee, "Event detection in twitter," in *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21*, 2011.

[39] B. West and P. Grigolini, *Complex Webs: Anticipating the Improbable*. Cambridge University Press, 2010.

[40] D. Zeng, H. Chen, R. Lusch, and S.-H. Li, "Social media analytics and intelligence," *Intelligent Systems, IEEE*, vol. 25, no. 6, pp. 13–16, 2010.

AUTHORS BIOGRAPHIES

Dr. Marco Furini has a Ph.D. in computer science and currently he is a faculty member of the Communication and Economics Department at the University of Modena and Reggio Emilia, Italy. His scientific interests include social computing and multimedia communication systems.

Dr. Simone Montangero has a Ph.D. in theoretical physics and he is Privat Dozent at the Institute for Quantum Information Processing at Ulm university, Germany. His scientific interests include quantum information theory, numerical methods and the study of complex systems.