Finding influential sources and breaking news in news media using graph analysis techniques

Iraklis Varlamis* and
Dimitrios Fassarakis Hilliard

Department of Informatics and Telematics,
Harokopio University of Athens,
9, Omirou St., Athens 17778, Greece
Email: varlamis@hua.gr
Email: it20935@hua.gr
*Corresponding author

Abstract: The popularity of social media has increased the interest for knowledge extraction from social networks and social media sites. The discovery of influential content or users and hidden social connections can be profitable for social media users and companies through personalisation and promotion respectively. Despite the abundance of works on social media and networks, there are no similar works in traditional (i.e., press, radio, TV) or online media (i.e., news sites). This work proposes a solution that solves the lack of influence or connection information by analysing news media content. Consequently, it detects the underlying influence among news media companies and provides knowledge about breaking news. Among the contributions of this work are: a new methodology for identifying and quantifying the implicit influence between news media, based on content similarity and a new method for the early detection of breaking news, with high interest to the mass media.

Keywords: news sites; influence; breaking news; graph analysis; news media analytics; blogosphere; social media analytics; trend analysis; prediction; web mining.

Reference to this paper should be made as follows: Varlamis, I. and Hilliard, D.F. (2017) 'Finding influential sources and breaking news in news media using graph analysis techniques', *Int. J. Web Engineering and Technology*, Vol. 12, No. 2, pp.143–164.

Biographical notes: Iraklis Varlamis is an Assistant Professor at the Department of Informatics and Telematics of Harokopio University of Athens. He received his PhD in Computer Science from Athens University of Economics and Business, Greece, and his MSc in Information Systems Engineering from UMIST UK. His research interests vary from data-mining and the use of semantics in web mining to social network analytics and knowledge extraction from social media and the news. He has published several articles in international journals and conferences, concerning web document clustering, the use of semantics in web link analysis and web usage mining, etc.

Dimitrios Fassarakis Hilliard graduated in the Department of Informatics and Telematics of Harokopio University of Athens in 2016. He has also worked for the Institute of Informatics and Telecommunication of the National Center for Scientific Research 'Demokritos' on research projects concerning the

Copyright © 2017 Inderscience Enterprises Ltd.

temporal evolution of social networks and has a publication in the International Conference on Web Intelligence Mining and Semantics (IEEE/WIC/ACM 2015). His research interests are on the fields of social network analytics and mining.

1 Introduction

The two main questions answered in this work with the analysis of news sites information are:

- Given a set of news sources that produce content in various topics, which are the most influential ones and which are the once that are mostly influenced by others?
- Is it possible to detect news that will quickly be reproduced by many sources and will soon become 'breaking news'?

An additional question that we answer in this work has to do with news content popularity and the possibility of a specific article to be reproduced by many other sources quickly and massively. With the term 'reproduce' we refer to identical content at a level higher than 90%. A text similarity algorithm at word level is used. Content reproduction considers only copies of an original content, which is quite common in news sites and blogs, and not referenced contented, e.g., through a hyperlink to the original content URL. This allows identifying breaking news that quickly spread across the news sphere without great variation in the textual content.

The only information that we use for answering the above questions is the content of the articles that the various news sources publish in their pages. We base our analysis on the simple assumption: "when two or more sources publish the same or almost identical content for an event but in different timestamps, then they are influenced by the first source that published the content".

In traditional newspapers, the team of editors usually creates new content for each event and publishes it as quickly as possible. In the case of blog-like news media, the majority of content is based on reposting the original content of online newspapers, and rarely original content is created. The current solution implements a news aggregator like Google News, which collects news from thousands of sources, online newspapers and blogs, and performs an automatic clustering of news items into highly coherent clusters. All documents in a cluster are highly similar and correspond to reproduction of the same original article. Using the article timestamp (related to the time the article has been crawled), we are able to identify the original source, the copiers and the delay of each reproduction. The implemented crawling mechanism allows checking the sources for updates every five minutes, which is a relatively small interval for creating identical content without copy. As a result, it is safer to rely to the article crawling timestamp, instead of the article editing timestamp, which can be edited by the user. If the article editing timestamp agrees with the crawling timestamp (i.e., is within the interval of the two consecutive crawls that found it for the first time) then we keep the editing timestamp, else we keep the crawling timestamp.

The solution has been designed to be integrated with an online aggregator that aims at picking the most important stories as they break, and featuring them as soon as they are available. In this paper we present our first findings from the analysis of the aggregated content for a period of two months, for news sites in Greece.

2 Related work

The idea of finding influential nodes in social networks has been well discussed in the literature, starting from the study of users' influence in social networks (Estévez et al., 2007; Microsoft, 2008; Trusov et al., 2010; Kale et al., 2007) and the ability of individuals to affect the dissemination of information (Kempe et al., 2003; Kimura et al., 2008; Kimura et al., 2007; Kim and Han, 2009) within a network. The research has been recently moved to social media (Nakajima et al., 2005; Song et al., 2007; Agarwal et al., 2008; Weng et al., 2010; Cha et al., 2010), where the aim is to model the influence of a user to other users and this is usually measured based on the interaction between a user and his/her followers (Almgren and Lee, 2015).

All the existing studies assume an explicit social network, where all users participate (network nodes), follow each other (network with directed edges) or form friendship bonds (network with undirected edges) and interact with each other, with actions that are predefined by the social media site (e.g., retweets, likes/dislikes, questions and replies). This explicitly declared information is used both for finding influential nodes and for locating content of high interest (e.g., influential posts) to the social network members.

For example, authors in (Cha et al., 2010) define three different activities that represent the different types of influence of a person in Twitter social network:

- *Indegree influence*, which counts the number of followers of a user and indicates the size of the audience for that user.
- *Retweet influence*, which counts the number of retweets containing a user's name and indicates the ability of that user to generate influential content.
- *Mention influence*, which counts the number of mentions containing a user's name and indicates the ability of that user to engage others in a conversation.

The analysis of 1.3 million Facebook users in Aral and Walker (2012), showed that influential users cluster together, whereas people that are susceptible to influence do not. This fact is important when seeking for influential users that can quickly 'spread the word' in the social network. In Halvey and Keane (2007) study the social interactions on YouTube and conclude that 'views' is the most popular type of interaction between users, compared to commenting, subscriptions, connection with friends, video uploading and sharing. In a similar context, Susarla et al. (2012) study the epidemics of video content in the social network formed in YouTube and result that subscribers are influenced first and the friends' network follows.

On a slightly different note, another common type of analysis is that of content ranking, in other words, finding 'influential' content, whether this is a product review, a blog or a tweet. Kritikopoulos et al. (2006), Adar et al. (2004) and Louta and Varlamis (2010) use the explicit structure of the blogosphere, with links and backward links in order to rank content and sources, to distinguish between spam and legitimate blogs, etc., in a manner similar to retweet and mention influence, which are defined above.

All the aforementioned works have several common characteristics that introduce new challenges for network analytics and related algorithms:

a The different type of edges and nodes that form the social network (i.e., implicit or explicit, permanent or temporal, positive or negative, directed or undirected when it comes to edges; users, items or groups when it comes to nodes).

b Multiple interconnected networks (e.g., a content sharing network bound to a social network,).

c The different type of structures of interest (i.e., cliques, influencers, outliers, trust and credibility, structure dynamics, structural holes, etc.).

Twitter, Facebook and the blogosphere are the most studied networks in terms of influence dynamics. Both positive and negative (e.g., spam) effects (Benevenuto et al., 2010; Gupta et al., 2014) have been studied in such environments. Several works in the literature, examine Twitter as a news reporting application. In Vis (2013) examines the use of Twitter as a reporting tool for breaking news, and compare the influence across mainstream or online media, twitterati, journalists or citizens actively engaged in the event. Similarly, in Kwak et al. (2010) analyse the follower-following topology of Twitter and identify influential users by topic. Despite the vast amount of research in measuring influence and content impact in social networks, the same problem in news media has not yet been studied. However, this is an issue of great importance for news media and online newspapers, since it is important for them to locate influential content and sources especially for companies that advertise products in online news media. Such companies want to know, which sources have major influence to the news-sphere and to their audience (Harada et al., 2015; Domingo et al., 2015). It is quite common in Greek news sites and blogs to reproduce (by copying the original content) an article published by a popular online newspaper, without an explicit reference to the source. This can be done without the newspapers consent or in agreement with the newspaper in order to increase the impact of the news (e.g., by reaching a local or thematic audience).

In their recent work Spitz and Gertz (2015) attempted to extract the citation network backbone of online news articles, which is based on explicit citations within the article context and resulted in a really sparse network. The result of this sparsity in explicit citations allowed authors to use only a handful of online newspapers and provide some basic analysis of the news media graph structure. According to our knowledge, our study is the first that analyses the social network that is implicitly formed among news media, by reproduction of content. The results from the application of our analysis to the Greek news sphere show that this reproduction is evident and is a strong criterion for determining whether a news article refers to a breaking event and whether a news source is influential.

Another important piece of knowledge for social media and their stakeholders is the early detection of breaking news. Breaking news is defined by Wiktionary[1] as "News that has either just happened or is currently happening. Breaking news articles may contain incomplete information, factual errors, or poor editing because of a rush to publication". Once again, Twitter and blogs have been studied as networks that can help in detecting breaking news (Phuvipadawat and Murata, 2010; Vis, 2013; Bruns et al., 2012). However, according to a study on the role of Twitter in news spreading (Subašić and Berendt, 2011) Twitter is mainly the tool for commenting on news and as such it cannot be considered neither a tool for creating nor for re-reporting existing news. As a consequence, social networks are ideal for measuring the impact of news and events to the public but it is more important to measure how fast news spreads among news media in order to distinguish between normal and breaking news at an early stage. Similarly in Macdonald et al. (2013) show that Twitter reports the same events as newswire providers, in addition to a long tail of minor events ignored by mainstream media. Concerning major news events, both streams indicate that the value that Twitter can bring in news setting comes predominantly from increased event coverage, not timeliness of reporting.

The proposed approach is not based on explicit links between news content, neither on explicit mentions among news sites. It rather collects news content from a wide range of news sources and clusters very similar content together. This allows us to early locate clusters that grow fast and thus discover breaking news in their early stage of creation. There is an intrinsic property in this type of 'breaking news', *newsworthiness*, which is more important in our case, since publishing an article or posting a blog requires an extra effort from the publishers side. Breaking news emerge frequently on the media, but only those that worth mentioning receive an increased amount of publicity, by being reproduced. The work of Shoemaker (2006), on news and newsworthiness provides a detailed discussion on the 'worthiness' property of news.

Our work has been designed in order to be included in an almost real-time setup, where new articles are collected every five minutes and updated information about existing clusters that grow or new clusters that appear are fed to our module in that pace.

3 Online news data

The aim of our analysis is to capture the dynamics and epidemics of online news sites and blogs using the similarity of published content as the main criterion for formulating a social graph between news sources. For this reason, it is important for the news crawler to collect as much information as possible concerning news content in Greek:

a from news sites that create original content

b from all possible news sources even those that simply reproduce original content from other sites.

Traditionally, news site crawlers collect information from the RSS feeds, which are provided by the news media. In the current study, we employ a news site crawler that is able to collect information both from RSS feeds and from the news web sites in the absence of RSS (Varlamis et al., 2014). Using this crawler, we are able to collect information from more than 1,400 news sites and 10,281 blogs. In the current work, we focus on the mechanism that analyses the articles, after they are collected and clustered. The methodology presented in this paper has been evaluated on a real dataset[2] comprising articles collected from Greek news sites and blogs during a two months period (February and March 2015) using the aforementioned crawling and clustering mechanism and has been served as a pilot for the service that we are deploying for a news aggregator portal in Greece, which aims in detecting 'breaking news' and promoting them to the portal's first page and also to find implicit links between the news media in Greece. The implicit links detected are then used to create the graph of news media influence in Greece, which can be useful to researchers that want to understand the media landscape in Greece. All the information is processed offline, but the approach has been designed in order to be incorporated to the online, real-time, environment.

3.1 Features

The information collected by the crawler comprises the title and content of the article, the publication date and time and the source of the article. Using a document clustering mechanism, we are able to create clusters of highly similar articles (≥90% similarity in content). The clustering algorithm has to be fast and incremental, since it operates in a stream-like process. It is a simple-pass clustering algorithm based on a text similarity threshold. When new articles are collected from the news crawler, it first checks if they can be assigned to any of the existing clusters (text classification). If the similarity between the article and any cluster is above a certain threshold, then the article is added to the closest cluster; otherwise it forms its own cluster. Since new articles are collected every few minutes, new clusters start with a small or medium number of articles and continue to grow depending on the interest of the topic across media. Usually, after a few hours or days the size of the cluster stabilises and a few days later clusters that do not grow any more are not further used in classification of new content.

We represent articles, clusters of articles and training articles per category in the Vector Space Model using *tf*/*idf* weighting at word level. Centroid clustering is applied when comparing an article to a cluster. Cosine similarity is the measure used for comparing articles (or cluster centroids). A document classification algorithm (a support vector machine classifier) is used for assigning clusters of articles to a broad thematic category (single-label). The thematic categories are predefined and correspond to the main categories covered by the press and online news media (i.e., politics, regional, sports, technology, etc.).

As a result the input for our news analysis mechanism comprises for each article the following features:

- Timestamp: the article's publication date and time.[3]
- Category: the thematic category that the article belongs to. Although it is reasonable that an article may span multiple thematic categories, for simplicity, we assume that there exists a dominant category for each article.
- Source: the news site or blog from which the article has been collected
- Cluster: the identifier of the cluster that the article belongs to.

Our algorithm processes each cluster separately and creates second-level information (e.g., by aggregating information from multiple clusters of the same topic) that will be used for the knowledge extraction tasks that follow (i.e., find breaking news and modelling news sites influence). The first step in this processing refers to the *date* field, which is used as a basis for measuring how fast an article is reproduced by other sources. Let us assume that cluster Cl_k comprises n articles $a_1 \ldots a_n$, each one published at the respective timestamp $t_1 \ldots t_n$. We can also assume that article a_j is the one that has been published earlier than any other article within cluster Cl_k at timestamp $t_{\min}$ ($t_j = t_{\min}$).

We define a new feature, which we call *publication delay* (d), for each article a_i in the cluster as follows:

$$d_i = t_i - t_{\min} \tag{1}$$

We also define the normalised publication delay, which we call *quickness* (q), for each article a_i in the cluster as follows:

$$q_i = 1 - \frac{t_i - t_{\min}}{t_{\max} - t_{\min}} \tag{2}$$

where $t_{\max}$ is the maximum timestamp value in the cluster.

According to equation (2) *quickness* is a number in [0…1] and denotes how quickly the original article is reproduced by another source. To give an example, let us assume that cluster has three articles a_1, a_2, a_3 with timestamps t_1, t_2, t_3 respectively, where $t_1 < t_2 < t_3$. Then $t_{\max} = t_3$, $t_{\min} = t_1$ and the quickness of a_1 is the maximum possible, $q_1 = 1$. Respectively $q_2 = 1 - \frac{(t_2 - t_1)}{(t_3 - t_1)}$ and $q_3 = 0$. The influence of a1 will be $i_1 = 2$, that of a_2 will be $i_2 = 1$ and that of a_3 will be $i_3 = 0$.

3.2 Influence levels

In a second processing step, information concerning the original article and its reproduction by other sites is examined in different aggregation levels. By this processing we are able to measure the aggregated influence of a source based on the originality of articles it publishes and the number of sources that reproduce its contents. For example, in a cluster that contains several similar articles from different sources only the source that published first will receive all the influence score. This process is repeated for all clusters and influence is aggregated in the following levels:

- Cluster influence: in this level, we measure the influence of a news source within the cluster. The cluster is a group of articles, retrieved from various news sources (i.e., sites and blogs), that have almost identical textual content. If the source publishes the first (i.e., minimum timestamp) article in the cluster then the influence will be equal to the number of articles in the cluster.
- Category influence: in this level, we focus on thematic categories and seek for the aggregated influence of a source in all clusters of the same category. For example, if the articles on 'politics' are grouped into N different clusters, and each cluster contains articles from various sources, we can measure an influence score for each source in each cluster and then aggregate the total influence of the source across all the N 'politics' clusters.
- Global influence: in this level we measure the overall influence of a source across all the news clusters. This allows us to identify the most influential news site in our dataset.

The aggregated influence $I(S_i)$ of a source i over the n clusters of the same level (i.e., same category or overall) is simply the sum of individual influence scores in each cluster:

$$I(S_i) = \sum_{k=1}^{n} I_k \,|\, (S_i), \forall S_i \in S \tag{3}$$

where S is the set of all sources we process, e.g., all the sources in the same cluster, or in the clusters of the same category and I_k is the influence of source S_i measured within the articles of cluster Cl_k.

Similarly, the aggregated quickness, which denotes how fast the news from a source is reproduced by other sources, is given by the equation (4):

$$Q(S_i)=\frac{\sum_{k=1}^{n} Q_k(S_i)}{n}, \forall S_i \in S \tag{4}$$

where $Q_k(S_i)$ is the quickness of source S_i measured within the articles of cluster Cl_k. If more than one articles from S_i appear in Cl_k, their average quickness is used.

Table 1 The sources in decreasing influence score order within a single cluster

CategoryId	*Original sourceId*	*SiteId*	*ClusterId*	*Influence*	*Quickness*
40	15	20602	12540656	205	0.874
80	15	20602	12518722	173	0.920
1	20	5055	12513458	137	0.824
1	20	645	12513458	115	0.836
1	20	21	12513458	90	0.840
2	19668	20602	12512829	87	0.956
1	20	4837	12513458	80	0.835
1	20	365	12513458	75	0.851
40	15	20602	12501287	74	0.978

Note: Such sites have been very influential for a specific event and also have quickness scores close to one.

Table 2 The influential sources by category

CategoryId	*Original sourceId*	*SiteId*	*Influence*	*Quickness*
64	25	2	955	0.999
1	13	9	899	0.999
2	4826	20195	693	0.997
64	2	25	635	0.999
1	13	558	526	0.997
64	8	405	478	0.993
64	8	4859	473	0.994
3	13	4859	425	0.992
40	15	20602	388	0.992

Table 3 The most influential sources overall original

Original sourceId	*SiteId*	*Influence*	*Quickness*
25	2	1249	0.9997
13	9	1041	0.9930
4826	20195	953	0.9861
13	558	857	0.9938
2	25	849	0.9990
16	4859	774	0.9951
9	558	681	0.9950
15	20602	580	0.9914
43	13	529	0.9870

The result from this process for the three aggregation levels that we have defined is similar to the one depicted in Tables 1 to 3. The columns in each table correspond to the category each cluster of articles belongs to (categoryId), the id of the site that published the first article in the cluster (original source id), the id of the site that published an article (say a_i) in the same cluster (siteID), the influence of this site (the number of articles in the cluster that have been published after a_i) and the quickness of the site based on the time that a_i has been published compared to all other articles in the cluster [using equation (2)]. The aggregation in Tables 2 and 3 is done using equations (3) and (4).

3.2.1 Dominant categories and influence

Finding the influential sources for every distinct category is a tedious and resource consuming task. In addition, the interest of readers and news agencies is mostly focused on a smaller set of categories, which we call *dominant* categories. *Dominant* are the categories that aggregate a significant amount of content (articles), which are frequently reproduced by other sources. As a result, the *dominant* categories are the ones with the maximum aggregated influence.

For this reason, we extend the definition of influence from a single source, to a category and define the aggregated influence of a category as follows: We assume a category C_j from the set of all categories C and every cluster Cl_k from the set of all article clusters that fall under category C_j. The aggregated influence of category C_j is defined as:

$$I(C_j) = \sum_{k=1}^{m} I(Cl_k), \forall \quad Cl_k \quad \text{classified in category} \quad C_j \tag{5}$$

where m is the number of articles categorised in C_j. The aggregated influence for a category according to equation (5) is the sum of sizes of all clusters that are classified (actually the articles within the clusters are classified) in category C_i.

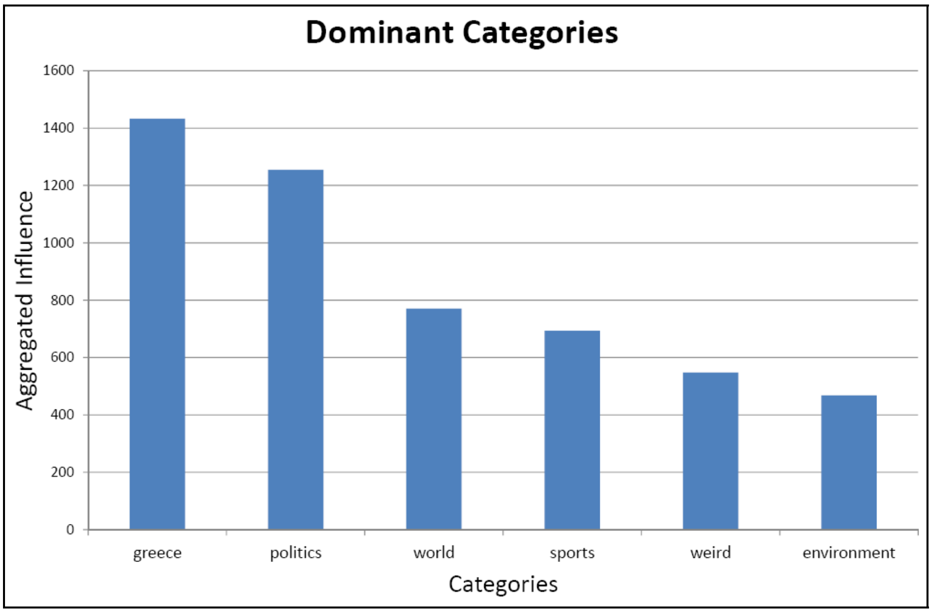
4 Experimental setup and results

4.1 *Finding the influencers and the influenced per category and overall*

4.1.1 *The dominant categories*

Using the methodology we defined in Section 3.2.1, we find the aggregated influence for all categories in the dataset we examined. The most influential categories (*dominant*) and their aggregated influence are depicted in Figure 1. From the figure we can see that the local news (category Greece) are the most influential among news sites and blogs, that frequently reproduce articles in this category. The results in Figure 1 seem to be in accordance with what one would expect to be the categories that attract readers' interest the most. What is less expected in these results is the category 'weird' Its presence can be explained by the tendency of blogs to reproduce weird news, which are published in other blogs, in order to increase their readability and attract more new readers.

Figure 1 A plot of the influence for the top-six most influential categories in our dataset (see online version for colours)



The next step of our analysis is to aggregate the overall influence across all categories and find the most influential sources. The results, presented in Figure 2 show that the sources that produce original content, which frequently is reproduced by other sites and blogs. Among the top-ten list we can see the websites of traditional newspapers that specialise in a domain (e.g., Naftemporiki in finance, Imerisia in politics, Protothema in news about Greece) and online media that also specialise in a topic (e.g., Euro2day in stock market) but also big blogs driven by well known journalists (e.g., enikos, zougla, thetoc, etc.).

Figure 2 A plots of the top-ten sources with the highest aggregated influence over all thematic categories in our dataset (see online version for colours)

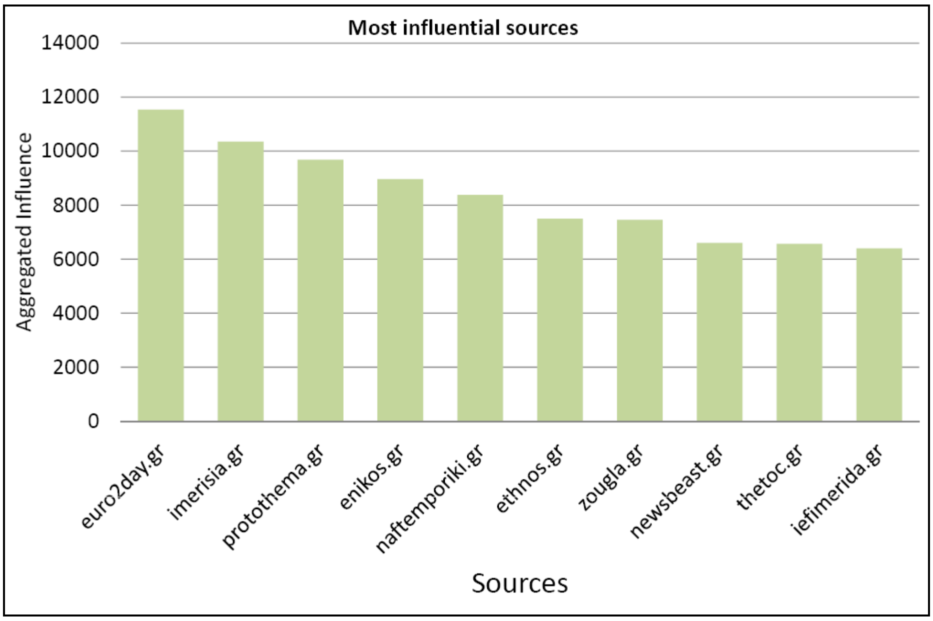
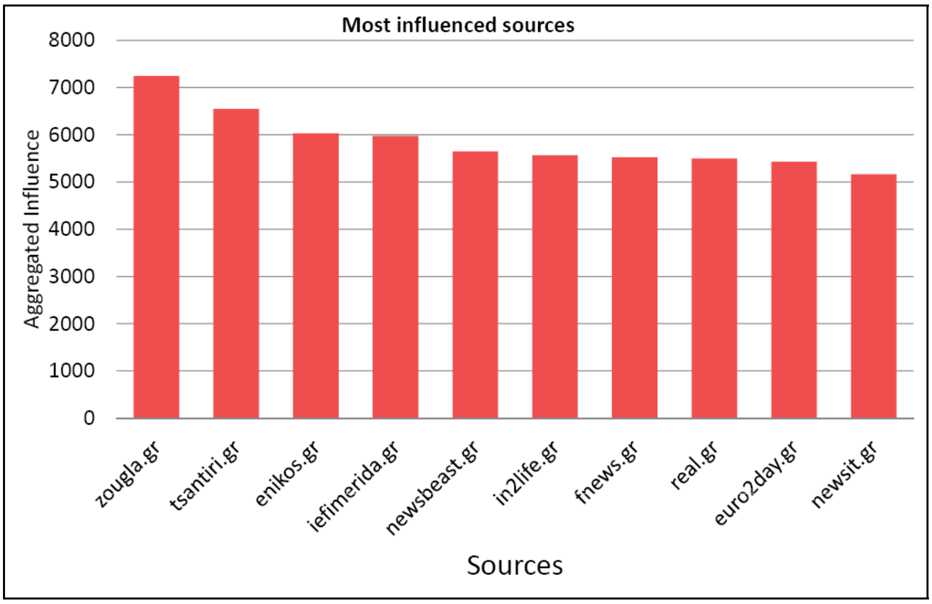
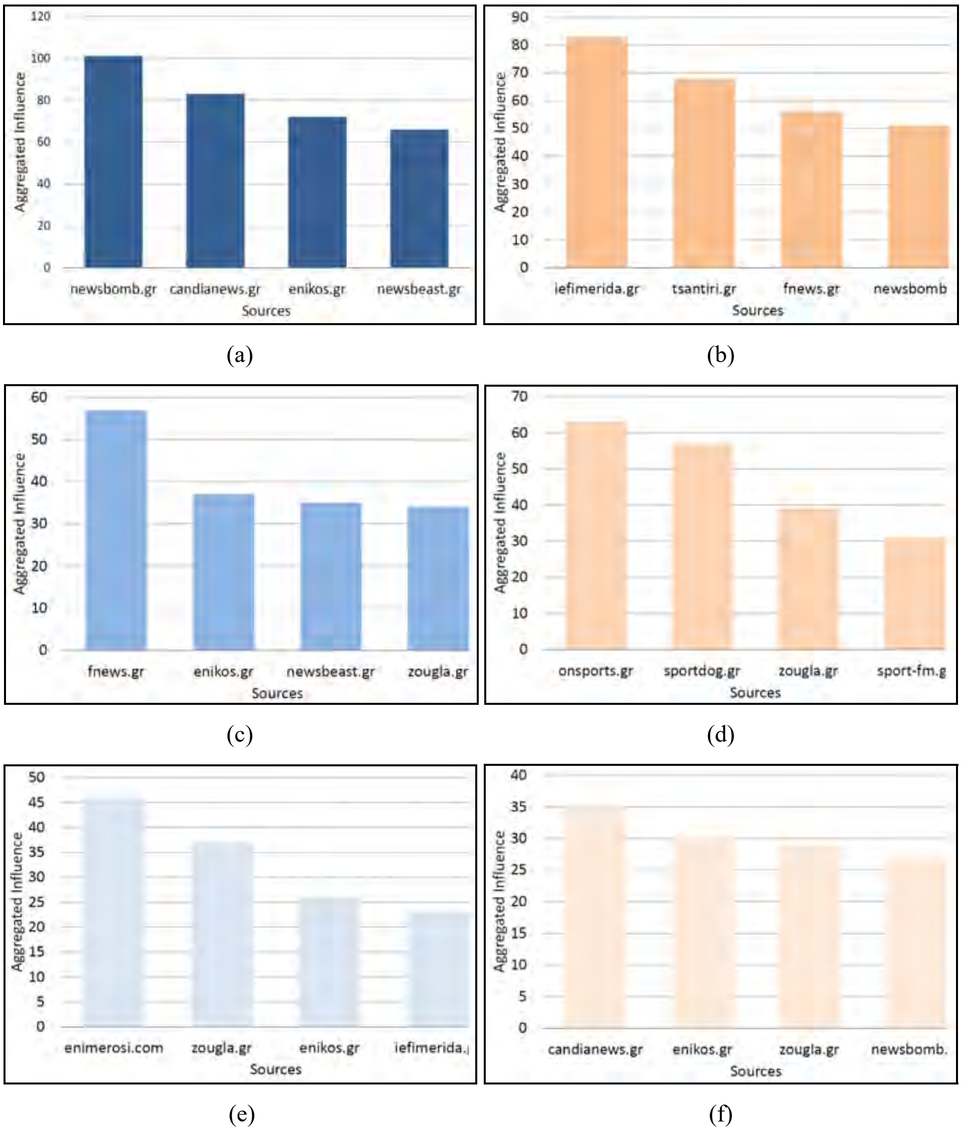


Figure 3 A plots of the top-ten sources that are influenced the most by other media across all thematic categories in our dataset (see online version for colours)



Next, we invert the concept of influence and focus on the sources that frequently reproduce others' article, thus spreading the news to more readers. The aggregated influence per source in this case is the sum of all articles that a certain source has reproduced from other sources. The results in Figure 3 present blog-like news sites that frequently use the content from other sources. An in-depth check shows that this is done officially with reference to the original source, especially from the sites in the top of our list. This is a way for this blog-like site to increase their content by ‘re-publishing’ content that has been created by their affiliated news media.

Figure 4 An example of the sources that have been mostly influenced by the most influential source in our dataset (euro2day.gr) for each of the dominant categories (a) category ‘Greece’ (b) category ‘politics’ (c) category ‘world’ (d) category ‘sports’ (e) category ‘weird’ (f) category ‘environment’ (see online version for colours)



(a) (b)

(c) (d)

(e) (f)

The next step is to analyse each category separately and find the influence to these sources aggregated per category. Using the most influential source as input (euro2day.gr), we aggregate all the reproductions of its articles from other sources across categories. The results of this analysis for the six dominant categories are depicted in Figure 4.

It is important to note here that among the most influenced sources are sources that still provide original content and influence others in some categories. However, they are not specialised in all categories, and they usually reproduce other sites news in the categories that fall out of their expertise. For example zougla.gr is a highly influential blog in category 'Greece', but usually reproduces content from other sites in 'sports' in order to satisfy its readers. The same holds for Sport-fm.gr, which is a 'sports' portal, but also reproduces content from other sources for all other categories. As a result, we see that there are news media that provide original content only, others that simply recycle existing content and others that do both.

4.1.2 Analysis of the influence graph

In this step, we examine the graph of influence between the various sources as a whole and not only in a node level. The graph has news sources as nodes and weighted directed edges that show which site is influenced by which other site and the degree of influence. Running the PageRank algorithm in this weighted and directed graph, we are able to find sources (news sites) of high centrality in this graph, sites that influence the most influential sites. BluePageRank centrality, accounts three distinct factors that determine the PageRank of a node (site in our case): PageRank centrality, which accounts three distinct factors that determine the PageRank of a node:

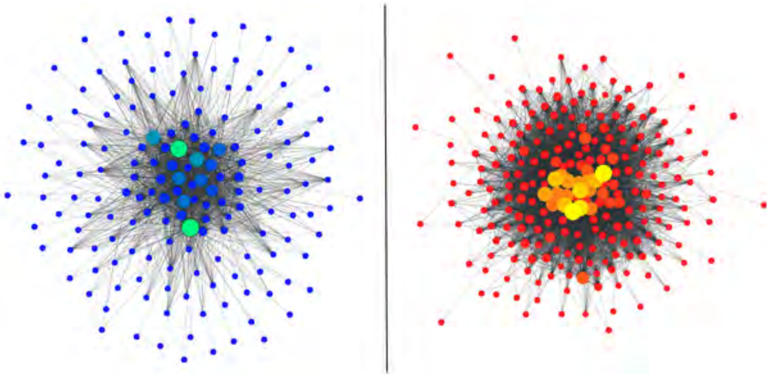
1 the number of links it receives (in degree)

2 the link propensity of the linkers (their outdegree),

3 the centrality of the linkers (links from important vertices are more valuable than those from obscure ones).

We create different graphs using the influence information only from the content of a specific category each time and create visualisations such as the ones depicted in Figure 5.

Figure 5 The graph of sources and influence edges for the category 'Greece' on the left and the complete graph for all the categories on the right (see online version for colours)



Note: The bigger nodes represent nodes with higher PageRank values.

Table 4 provides the news sources with the highest PageRank score in the influence graph. The results are in accordance to those depicted in the influential sources figure (Figure 2). However, a more careful examinations shows that sources such as *zougla:gr* or *sport-fm:gr*, rank higher in this table, and this is because they influence the media sphere by publishing novel, topic specific content, but also reproduce content from other sources. Taking into account results from other researches in Greek online media[4], we see that the news portals of Table 4 are among the top most visited sites in Greece.[5]

Table 4 The news media with the highest PageRank score in the influence graph

PageRank values	
News source	*Value*
Euro2day.gr	0.0478238661685
Imerisia.gr	0.0470638430403
Protothema.gr	0.0426127786659
Enikos.gr	0.0395576157629
Zougla.gr	0.036728189514
Naftemporiki.gr	0.0362524003234
Ethnos.gr	0.0336622515437
Sport-fm.gr	0.0308966813659
Thetoc.gr	0.029691758954
Newsbeast.gr	0.0285247409145

4.2 Identification of breaking news

The next step of our analysis is to create a machine learning model for detecting ‘breaking news’ at an early stage. Since our analysis is based on clusters of articles that are very similar in content, our methodology will be based on the identification of those features that have the highest predictive value for the clusters of articles.

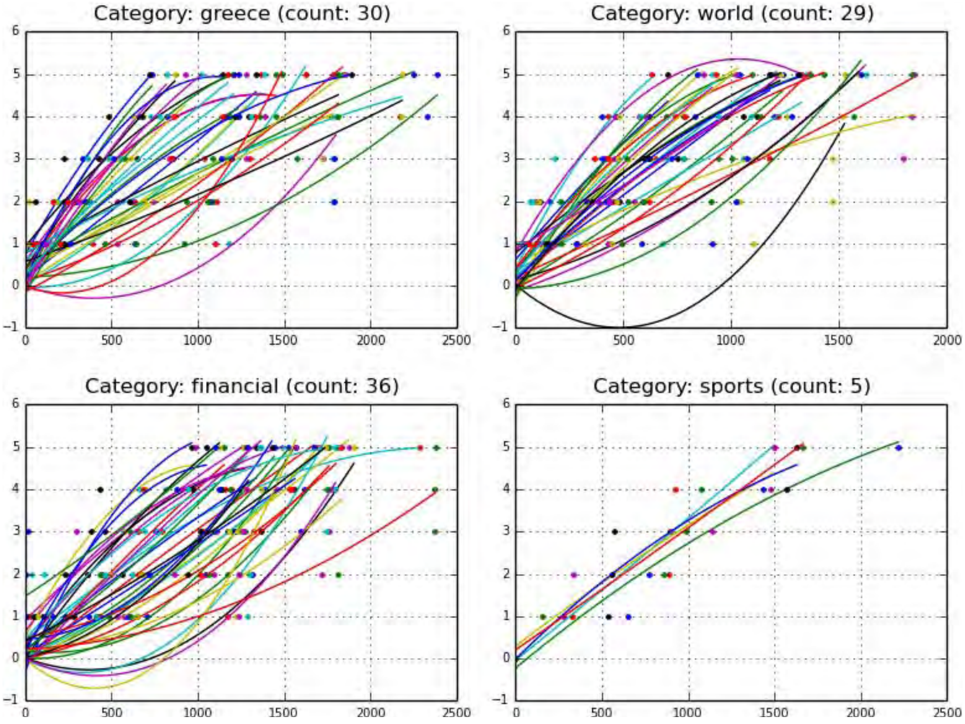
Our main aim is to distinguish between clusters of articles that discuss a topic of highly increasing interest and clusters that are of average or low interest which remains stable over time. As a result, we are interested in clusters that begin with a high number of articles when they are created and continue to grow fast. Since clusters are updated periodically it is possible that two clusters grow equally between two consecutive check points, but the distribution of articles during this period differs significantly. So it is important to examine the exact time of publication, for each article in the cluster. Based on the above, the features that seem to be most appropriate for distinguishing ‘breaking news’ are:

- the size of the cluster
- the rate of its growth.

In order to improve the time performance of our algorithm, we ignore very small clusters (i.e., clusters with size less than 15). We refer to this threshold onwards as *influence_threshold*. Concerning the rate of a cluster growth and the detailed information of when each article within a cluster was published, we transform the original *publication_time* feature to a new feature called *delay*, which represents the seconds between an article's publication time and the publication time of the first article in the cluster. Using *delay* as feature, we have a set of delay values for each cluster (one value for each article in the cluster). Since we are interested in detecting ‘breaking news’ as early as possible, it is important to examine how many delay values (we call them *delay_points* onwards) we need to determine whether a cluster corresponds to a ‘breaking news’ cluster. For this purpose, we fit a curve on the set of the first N *delay_points* used for each cluster (in our analysis $N = 5$). Since news in different categories are reproduced with different rates, it is important to train different models for each news category. We use a training data set that comprises clusters that do not grow any more, which are characterised as ‘breaking’ or ‘non-breaking news’ after examining their final size and the duration of their lifeline (for how long they were growing in size). We select clusters that clearly differ between the two sets and plot their growth curves. Figures 6 and 7 that follow, depict the curves for the ‘breaking news’ and ‘non-breaking news’ clusters in our dataset, for four different categories. We fit all curves using the first five *delay_points* of each cluster and a second degree polynomial (i.e., $y = ax^2 + bx + c$).

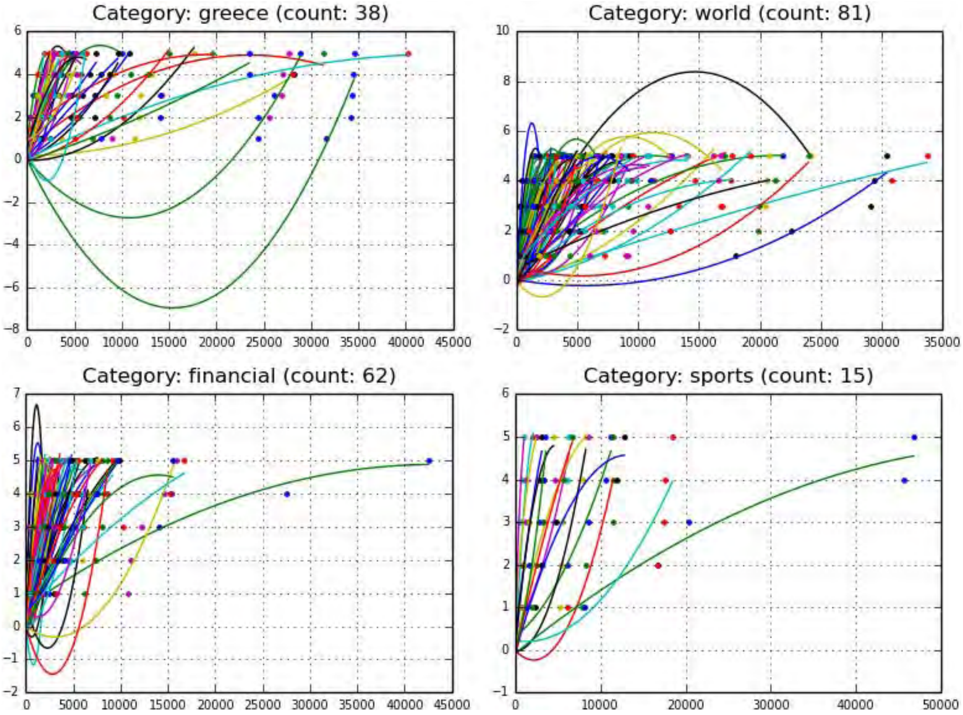
The comparison between the two figures clearly depicts the differences between ‘breaking news’ clusters that have grown fast, reaching size five in the first 20 minutes and ‘non-breaking news’ clusters that need several hours to reach the size of five.

Figure 6 The curves that fit the growth of each ‘breaking news’ cluster using the first five *delay_points* (see online version for colours)



Notes: The delay is depicted in the horizontal axis in seconds. The vertical axis contains the size of each cluster. The *count* = *N* above each chart represents the number of clusters in each category.

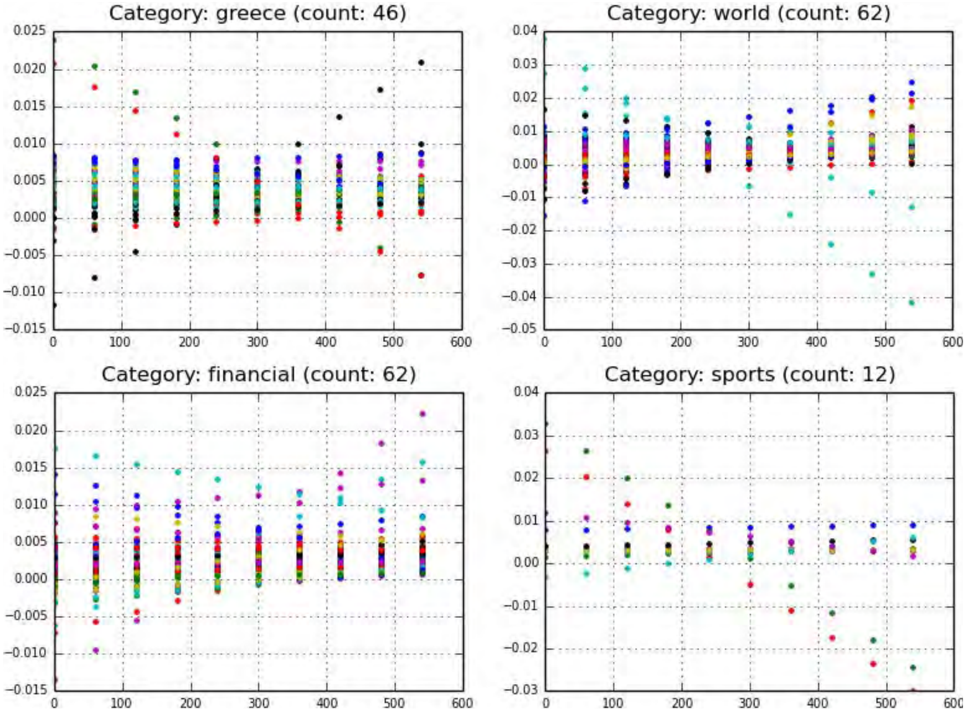
Figure 7 The curves that fit the growth of each ‘non-breaking news’ cluster using the first five *delay_points* (see online version for colours)



Notes: The delay is depicted in the horizontal axis in seconds. The vertical axis contains the size of each cluster.

Based the above analysis, we are able to model the relation between the size of a cluster and time, so we need to define the time step (*Time_period*) between consecutive checks, in order to better determine ‘breaking news’ clusters. We compute the slope of each curve depicted at Figures 6 and 7 at different time points (every one minute) for the first ten minutes of the cluster life. The results are depicted at Figures 8 and 9. Although there are variations among the different categories and among clusters, most of the ‘breaking news’ clusters have a positive slope during the first ten minutes, which usually grows after the first five minutes, whereas ‘non-breaking news’ clusters even if they start with a positive growth rate, they usually reach a zero slope (stop growing) at the first ten minutes. These ten values (the slope of the curve in the first ten minutes, measured every minute), are the ten features that our machine learning algorithm uses as input in order to decide whether a cluster corresponds to ‘breaking news’.

Figure 8 The evolution of slopes for the ‘breaking news’ clusters’ curves during the first ten minutes (600 seconds) (see online version for colours)

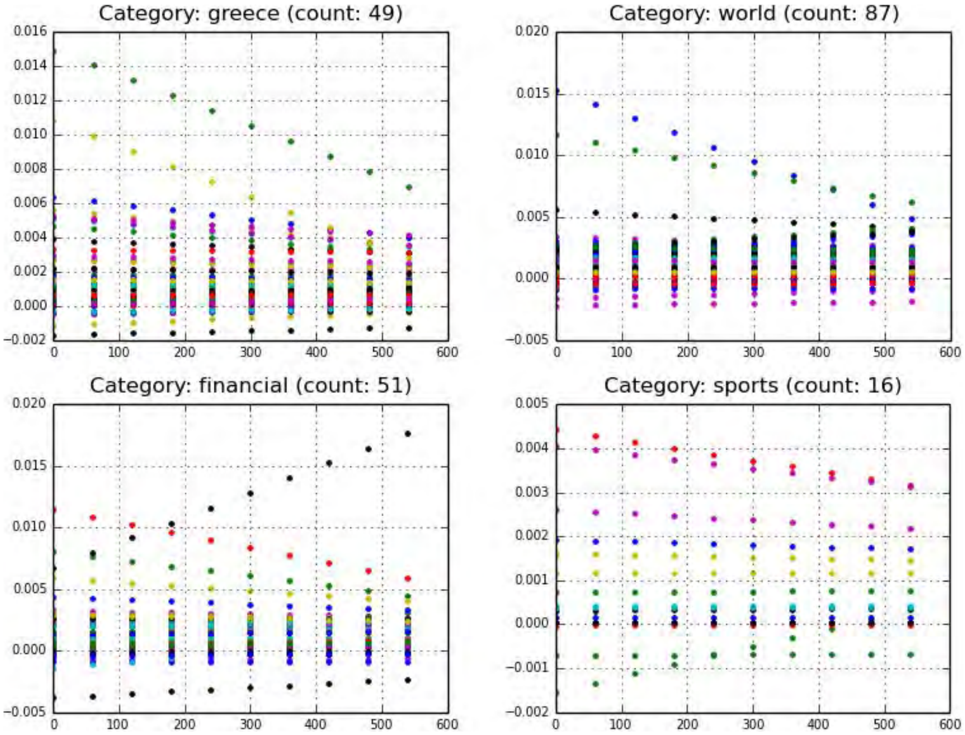


Note: The vertical axis contains the slope value and the horizontal axis the time in seconds.

The classifier that we employed for training our models was random forest (Liaw and Wiener, 2002) and its implementation in scikit-learn.[6] We used ten estimators (number of trees created using a random feature subset), which is the only parameter of Random Forest. We evaluated the performance of the Random Forest algorithm using the first five *delay_points* and three different types of curve fit (first, second and third polynomial, deg = 1, 2, 3 respectively) and pruned out small clusters (size < 15). We performed 10-fold cross validation on a set of clusters that we labelled as ‘breaking’ and ‘non-breaking’ using their final size and their period of growth[7], similarly to what we did for the training set. In Table 5 we present the Precision, Recall and F-measure scores for each class separately and overall.

Using the second polynomial curve for fitting the points[8], we measure the effect of the number of *delay_points* to the algorithm performance. We take an increasing number of *delay_points*, repeat the whole process and measure the overall $F_1 - score$.[9] The results of this process are depicted in Figure 10 and show that using 10–12 *delay_points* can increase the prediction performance, which reaches 0.97 in our case. However, using more points reduces the quality of results, which means that there is no reason to monitor the cluster growth for more than 15 minutes in order to predict whether it will be a breaking-news cluster or not.

Figure 9 The evolution of slopes for the 'non-breaking news' clusters' curves during the first ten minutes (600 seconds) (see online version for colours)

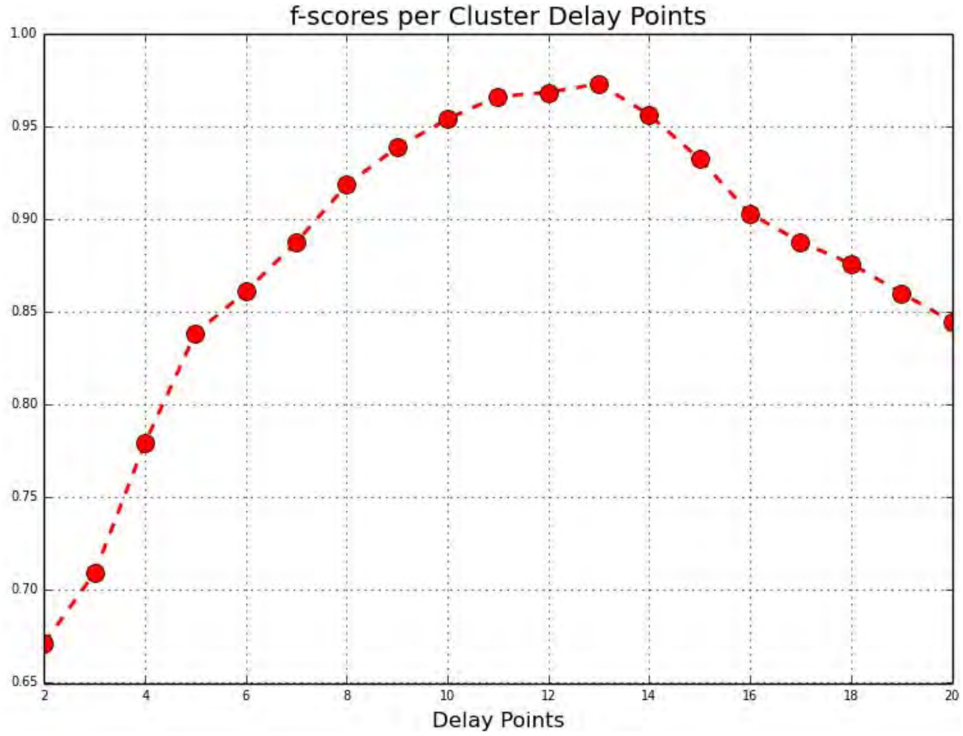


Note: The vertical axis contains the slope value and the horizontal axis the time in seconds.

Table 5 The performance of distinguishing between breaking and non-breaking news using the first five *delay_points* to fit a curve and the projected numbers in the first ten minutes

Parameters: deg =1, Influence threshold = 15, delay points = 5				
Classifier scores	*Precision*	*Recall*	*F-measure*	*Number of clusters*
Breaking set	0.83	0.81	0.82	423
Non-breaking set	0.81	0.83	0.82	412
Overall	0.82	0.82	0.82	835
Parameters: deg =2, Influence threshold = 15, delay points = 5				
Classifier scores	*Precision*	*Recall*	*F-measure*	*Number of clusters*
Breaking set	0.86	0.83	0.85	423
Non-breaking set	0.82	0.85	0.84	412
Overall	0.84	0.84	0.84	835
Parameters: deg =3, Influence threshold = 15, delay points = 5				
Classifier scores	*Precision*	*Recall*	*F-measure*	*Number of clusters*
Breaking set	0.83	0.82	0.83	423
Non-breaking set	0.82	0.83	0.83	412
Overall	0.83	0.83	0.83	835

Figure 10 The performance of our prediction model for different number of *delay_points* (see online version for colours)



As far as it concerns the time complexity and scalability of the approach, it is feasible to apply it to a real-time setup, since the main bottleneck is the crawling process, which in the current setup takes place every five minutes. Once the classification model is trained for a category, using historical data about breaking and non-breaking news' clusters, it takes less than a second to get a prediction for hundreds of recently created clusters.

5 Conclusions

This work models the influence among news media (sites and blogs) using only the similarity between content that the different media publish online. It is based on a simple assumption that when a news site or blog posts a new article, which is highly similar to another article published earlier in another site, then there exists a strong, but implicit, influence between the two. Using the proposed text similarity and text clustering approach, we are able to:

a Create the graph of news media influence, where news sites are nodes and edges correspond to a large number of similar posts between two sites.

b model the growth of clusters that correspond to different news and detect ‘breaking news’ clusters in their early stage, when only a handful of sources have only reproduced the article contents.

Although in the analysis of influence, we focused on the detection of highly influential sources, it is on our next plans to study the inter-organisational influence, quantify the ties that exist between news media companies and monitor their evolution.

The methodology presented in this paper has been evaluated on a real dataset comprising articles collected from Greek news sites and blogs during a two months period (September and October 2015) and has been served as a pilot for the service that we are deploying for a news aggregator portal in Greece, which aims in detecting 'breaking news' and promoting them to the portal's first page and also to find implicit links between the news media in Greece. The next steps of our work comprise the analysis of the influence graph and the detection of more interesting structures, such as cliques, hubs and authorities in the Greek online media sphere.

References

Adar, E., Zhang, L., Adamic, L. and Lukose, R. (2004) *Implicit Structure and the Dynamics of Blogspace*, In Workshop on the Blogging Ecosystem, WWW.

Agarwal, N., Liu, H., Tang, L. and Yu, P.S. (2008) 'Identifying the influential bloggers in a community', in *ACM WSDM*, 2008.

Almgren, K. and Lee, J. (2015) 'Who influences whom: content-based approach for predicting influential users in social networks', in *International Conference on Advances in Big Data Analytics*, pp.89–99.

Aral, S. and Walker, D. (2012) 'Identifying influential and susceptible members of social networks', *Science*, Vol. 337, No. 6092, pp.337–341.

Benevenuto, F., Magno, G., Rodrigues, T. and Almeida, V. (2010) 'Detecting spammers on Twitter', in *Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS)*, Vol. 6, p. 12.

Bruns, A., Highfield, T. and Lind, R.A. (2012) 'Blogs, Twitter, and breaking news: the produsage of citizen journalism', *Produsing Theory in a Digital World: the Intersection of Audiences and Production in Contemporary Theory*, Vol. 80, No. 2012, pp.15–32.

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, P.K. (2010) 'Measuring user influence in Twitter: the million follower fallacy', *ICWSM*, Vol. 10, Nos. 10–17, p.30.

Domingo, D., Masip, P. and Costera Meijer, I. (2015) 'Tracing digital news networks: towards an integrated framework of the dynamics of news production, circulation and use', *Digital Journalism*, Vol. 3, No. 1, pp.53–67.

Estévez, P.A., Vera, P.A. and Saito, K. (2007) 'Selecting the most influential nodes in social networks', in the *International Joint Conference on Neural Networks, IJCNN*.

Gupta, A., Kumaraguru, P., Castillo, C. and Meier, P. (2014) 'Tweetcred: real-time credibility assessment of content on Twitter', in *International Conference on Social Informatics*, Springer, pp.228–243.

Halvey, M.J. and Keane, M.T. (2007) 'Exploring social dynamics in online media sharing', in *Proceedings of the 16th International Conference on World Wide Web*, ACM, pp.1273–1274.

Harada, J., Darmon, D., Girvan, M. and Rand, W. (2015) 'Forecasting high tide: predicting times of elevated activity in online social media', in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ACM, pp.504–507.

Kale, A. Karandikar, A., Kolari, P., Java, A., Finin, T. and Joshi, A. (2007) 'Modeling trust and influence in the blogosphere using link polarity', in *ICWSM*.

Kempe, D., Kleinberg, J.M., and Tardos, È. (2003) 'Maximizing the spread of influence through a social network', in *KDD*, pp.137–146.

Kim, E.S. and Han, S.S. (2009) 'An analytical way to find influencers on social networks and validate their effects in disseminating social games', in *ASONAM*, pp.41–46.

Kimura, M., Saito, K. and Nakano, R. (2007) 'Extracting influential nodes for information diffusion on a social network', in *AAAI*, pp.1371–1376.

Kimura, M., Yamakawa, K., Saito, K. and Motoda, H. (2008) 'Community analysis of influential nodes for information diffusion on a social network', in *IJCNN*, pp.1358–1363.

Kritikopoulos, A., Sideri, M. and Varlamis, I. (2006) 'Blogrank: ranking weblogs based on connectivity and similarity features', in *Proceedings of the 2nd International Workshop on Advanced Architectures and Algorithms for Internet Delivery and Applications*, ACM, p. 8.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010) 'What is twitter, a social network or a news media?', in *Proceedings of the 19th International Conference on World Wide Web*, ACM, pp.591–600.

Liaw, A. and Wiener, M. (2002) 'Classification and regression by random forest', *R News*, Vol. 2, No. 3, pp.18–22.

Louta, M. and Varlamis, I. (2010) 'Blog rating as an iterative collaborative process', in *Semantics in Adaptive and Personalized Services*, Springer, pp.187–203.

Macdonald, C., McCreadie, R., Osborne, M., Ounis, I., Petrovic, S. and Shrimpton, L. (2013) 'Can Twitter replace newswire for breaking news', in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.*

Microsoft (2008) 'Identifying influential persons in a social network', *US Patent Application*, #20080070209.

Nakajima, S., Tatemura, J., Hino, Y., Hara, Y. and Tanaka, K. (2005) 'Discovering important bloggers based on analyzing blog threads', in *2nd Annual Workshop on the Blogging Ecosystem: Aggregation, Analysis and Dynamics.*

Phuvipadawat, S. and Murata, T. (2010) 'Breaking news detection and tracking in Twitter', in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, IEEE, Vol. 3, pp.120–123.

Shoemaker, P.J. (2006) 'News and newsworthiness: a commentary', *Communications*, Vol. 31, No. 1, pp.105–111.

Song, X., Chi, Y., Hino, K. and Tseng, B. (2007) 'Identifying opinion leaders in the blogosphere', in *CIKM*, pp.971–974.

Spitz, A. and Gertz, M. (2015) 'Breaking the news: extracting the sparse citation network backbone of online news articles', in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ACM, pp.274–279.

Subašić, I. and Berendt, B. (2011) 'Peddling or creating? investigating the role of Twitter in news reporting', in *Advances in Information Retrieval*, Springer, pp.207–213.

Susarla, A., Oh, J-H. and Tan, Y. (2012) 'Social networks and the diffusion of user-generated content: evidence from YouTube', *Information Systems Research*, Vol. 23, No. 1, pp.23–41.

Trusov, M., Bodapati, A.V. and Bucklin, R.E. (2010) 'Determining influential users in internet social networks', *Journal of Marketing Research*, Vol. 47, No. 4, pp.643–658.

Varlamis, I., Tsirakis, N., Poulopoulos, V. and Tsantilas, v (2014) 'An automatic wrapper generation process for large-scale crawling of news websites',in *Proceedings of the 18th Panhellenic Conference on Informatics*, ACM, pp.1–6.

Vis, F. (2013) 'Twitter as a reporting tool for breaking news: journalists tweeting the 2011 UK riots', *Digital Journalism*, Vol. 1, No. 1, pp.27–47.

Weng, J., Lim, E-P., Jiang, J. and He, Q. (2010) 'Twitter rank: finding topic-sensitive influential twitterers', in *WSDM*, pp.261–270.

Notes

[1] Https://en.wiktionary.org/wiki/breaking news

[2] The dataset can be downloaded from https://www.dit.hua.gr/ varlamis/datasets/

[3] For simplicity we refer to publication timestamp extracted from the article itself, although we can use the timestamp of the crawling which cannot be biased.

[4] Maria Kontochristou, Nagia Mentzi, Media Landscape in Greece [online] http://ejc.net/medialandscapes=greece.

[5] From the study of Kontochristou and Mentzi: "According to Alexa, a web information company, among the top sites users visit in Greece are: in.gr (ninth overall), zougla.gr (18th most popular), naftemporiki.gr (28th most popular)".

[6] Http://scikit-learn.org/

[7] Two human annotators manually classified 900 clusters to 'breaking' or 'non-breaking' and the inter-annotator agreement ratio was slightly above 90%. We kept only the clusters where the two annotators agreed for the evaluation.

[8] Since the crawler updates the information for each cluster (size) every five minutes, we use curve fitting in order to project the cluster size in the intermediates.

[9] $F_1 = \frac{2 * precsision * recall}{precission + recall}$