Detecting anomalies in open source information diffusion

François Nel, Antoine Carré, Philippe Capet, Thomas Delavallade

Detecting anomalies in open source information diffusion

François Nel
LIP6 -Université Pierre et Marie Curie-Paris6, UMR7606,
104, avenue du président Kennedy, Paris, F-75016 France

francois.nel@lip6.fr

Antoine Carré
École Polytechnique
Route de Saclay, 91128 Palaiseau Cedex

antoine.carre@polytechnique.edu

Philippe Capet
Thales Land and Joint Systems
160, boulevard de Valmy - BP 82 - 92704 Colombes Cedex - France

philippe.capet@fr.thalesgroup.com

Thomas Delavallade
Thales Land and Joint Systems
160, boulevard de Valmy - BP 82 - 92704 Colombes Cedex - France

thomas.delavallade@fr.thalesgroup.com

ABSTRACT

The adoption of the Internet as a massive information diffusion medium has considerably modified information dynamics. An increasing amount of available information from uncontrolled, various and unreferenced sources makes this new mediatic environment suitable for various information diffusion phenomena like amplification phenomena that may affect significantly political, strategical or economical matters.

Strategical or economical intelligence analysts using open sources have to adapt their methods to face significant and changing information flows. In this context, they need automatic tools to select interesting phenomena among large quantities of data.

This paper focuses on substantial variations in open source information diffusion that we call anomalies. Firstly, we introduce a model of a website network considering relevant parameters taking part in the dynamics of information flows over the Web. Then we propose a methodology based on this network to detect anomalies.

1.0 INTRODUCTION

When a great number of people suddenly focuses on a topic and consistently comments on it, it becomes a potential target and increases its mediatic vulnerability. In many cases this phenomenon is virulent, sudden and comes with over mediatic coverage under the pressure of which leaders and decision makers are conducted to react in a hurry.

Nowadays, the Internet plays a major role in information diffusion. More important, its adoption has changed the way of spreading information by giving everyone not only the possibility to observe information flows but also the opportunity to influence and create them.

By charting how often a particular search term is entered, search engines like Google trends [1] are able to evaluate people themes of interest at a given time which could be interpreted as a biased *a posteriori* detection of *amplification* phenomena. These mediatic phenomena are not limited to insignificant subjects: they can have major consequences on political, strategical or economical decisions. Increasingly, they are triggered on purpose for various reasons: campaigns can be carried out in order to discredit a company, endanger strategical choices or question political decisions.

Nowadays, many sources or intermediaries are available in open source. Information websites based classically on the model of traditional media are now mixed with autonomous and personal publishing modalities. Forums, chatrooms and above all weblogs are rapidly growing means to create and spread information. These new influential intermediaries are playing a major share in the quickening and amplification of information diffusion and consequently they have to be considered in any study of information dynamics on the Web.

The main goal of this paper is to study in details the components of a system designed to detect major variations in information diffusion that we call *anomalies*. These variations are relevant symptoms of more complex underlying phenomena and their automatic detection is very useful for economical or strategical intelligence analysts. We base the system on a model of source network containing relevant parameters on information diffusion. Basically, the anomaly detection process consists in comparing a reference network with the one to be analysed and in selecting major variations.

The paper is organised as follows: in section 2, we first depict an overview of related research to expose the needs and novelties of this study. In section 3, we introduce a network of source containing information diffusion data. Section 4 presents the methodology used to build a source network and section 5 exposes the anomaly detection process. Lastly, section 6 concludes and presents forthcoming works.

2.0 RELATED WORK

Information propagation is related to the more common issue of contagion in its general definition. The introduction of mathematical models of contagion has been inspired by researches conducted in social and biological sciences.

In biological sciences, the SIR model [4] is classically used to represent the mechanisms of epidemic propagation: firstly, a person is *susceptible* to contract a disease, and then if exposed there is a probability that she may be *infected*. Finally after a certain period of time, she *recovers*. The behavior of a blogger on the Web may be interpreted using the SIR model. Indeed after being exposed to a topic read on a website, a blogger may decide to write an article about it on its own blog/website. As long as her new article is available to readers, she is infected as she takes part in the diffusion of the topic over the Web. She recovers the moment the article is no longer available, or when the page is no longer visited.

Despite the diversity of application areas, models of contagion are all grouped into two categories [5]: *independent interactions models* and *threshold models*. These categories differ in the conditions for an individual to go from a state of *susceptible* to a state of *infected*.

In independent interactions models, there is no interdependency between successive exposures. In other words, when exposed, an individual will have the same probability to be infected whatever her history of exposure. The SIR model exposed above is an example of an independent interaction model. Some models of social contagion like the *independent cascade model* also fall into this category.

On the contrary, threshold models suppose that the infection occurs when a critical number of exposures has been reached. These models are historically used for social contagion where an individual adoption of a belief is influenced by the conviction of all her acquaintances.

Recently, some research has been conducted in order to generalize the existing models and their variations. Motivated by marketing issues, the *general cascade model* [8] sets aside the independence assumption of the *independent cascade model*. Similarly, in biological sciences, [4,5] introduce a generalized model of contagion that interpolates independent interactions models and threshold models.

A major motivation in simulating the spread of information in social networks is to identify a set of influential nodes in order to maximize exposure [8,9]. Moreover, [14] reveals the importance of the hierarchical structure of the network in the spread dynamics.

In the case of information diffusion on the Web and more precisely on the blogosphere, classical models like the threshold model [7] or the independent cascade model [6] have often been used. Nevertheless, many influential factors particular to this mode of social diffusion have not been taken into account like for instance the area of interest of each website or the way information is presented. In this connection, the blogosphere is a very promising area of research and it is worth mentioning that recently an evaluation campaign TREC [11] has been dedicated to this subject.

This paper intends to properly define the most relevant parameters that should be included in a system to model information dynamics and detect anomalies on the Web. Notably, it gives a major importance to the network structure and to the areas of interest and influence of websites.

3.0 PROPOSED MODEL

Ideally, a model of information diffusion on the Web should be able to simulate the state of awareness of Web users about a specific piece of information. Nevertheless, from a practical point of view it seems difficult to have access to such data. Moreover, in such a study, only Web users who may have an impact on information diffusion (in other words who routinely publish information on a website) could be taken into account.

Therefore, the proposed model does not consider Web users individually but will only focus on websites publications. The global awareness will be represented by the way information is depicted on the Web and how the readers have access to it.

The model is based on a network of websites. Each node of the network represents a website. Let $W = \{W_i, i \in I\}$ be the set of all nodes of the network. The nodes are linked to each other by directed edges meaning *"is a source of information for"*. This way, information propagation can be easily monitored following the directed edges on the network.

As it is impossible to know the sources of information of every person publishing on the Web, we base our representation of the source network on the following hypothesis: when a publisher explicitly refers to another Web page using a hyperlink, the website pointed by the cited link is considered as one of the information sources for the publisher.

3.1   Expressing the importance of a link

Let $M = [m_{i,j}]_{n \times n}$ be the weighted adjacency matrix of the network of websites ($n = |I|$). The term $m_{i,j}$ expresses the value of the edge from W_i to W_j. Practically, this edge represents a set of hyperlinks pointing to W_i cited by W_j. We want that its value reflects how the citation relation of W_i on W_j is visible on the

Web. For each hyperlink, this value increases with the importance of the hyperlink. For that, we choose to define it as the sum of the weights of its hyperlinks.

In the following, four different methods to establish how to value an hyperlink are described. They are presented in the order of growing complexity, some are global as they only consider the website publishing the link whereas some also take into account the article containing the link.

• **Basic evaluation**

In this basic evaluation, every hyperlink is given the same value 1 without further consideration. Thus, the term $m_{i,j}$ of the weighted adjacency matrix is the number of hyperlinks to W_i cited by W_j.

• **Absolute influence of a website**

We consider in this method, that the importance of a piece of information published on a website depends on the influence of the website publishing this information on the Web.

If available, standard statistics of visitors activity may be used to evaluate this influence degree. Otherwise, the results of a PageRank algorithm [12] on the set W could be used. The InfluenceRank [13] which takes into account the information novelty to detect opinion leaders in the blogosphere could be another option.

We define the value of every hyperlink cited by W_j as its absolute influence degree noted λ_j. The previous method consists in setting $\lambda_j = 1$.

• **Using categories**

Many websites and particularly weblogs are very specialized. It means that information published on the Web goes through a very selective process which depends on the area of interest of the website and the topic of the published information.

Therefore we introduce a set of categories $C = \{C_k, k \in K\}$ of interest and an influence degree in each category of interest for each website. Let $\lambda_{j,k}$ be the influence degree of W_j website in the category C_k. $\lambda_j = (\lambda_{j,k})_{k \in K}$ is called the *influence vector* of W_j.

Just like a website is linked with categories of interest by its influence degree λ, an article has a membership degree to each category of interest.

Let $\mu_{r,k}$ be the membership degree of article A_r to category C_k. $\mu_r = (\mu_{r,k})_{k \in K}$ is called the *membership vector* of A_r.

The value of every hyperlink cited by W_j in article A_r is now defined as a *compatibility degree Comp* between the vectors λ_j and μ_r, $Comp(\lambda_{j,}\ \mu_r)$.

An example of compatibility measure could be choosing Comp as a weighted dot product,

$$Comp(\lambda_j, \mu_r) = \sum_{k \in K} \alpha_k \lambda_{j,k} \mu_{r,k}$$

where the factor α_k would represent the relative importance of the category k.

- **Using article visibility**

The last parameter we introduce to model an article is named *visibility* in this paper. It measures the importance a website gives to an article as a function of how the website displays it and how a reader will have access to it.

Let $v_{r,j}$ be the visibility of article A_r on website W_j. The following parameters can be considered to evaluate it:

- The *size* of the article. Basically it is the normalized number of words of the article (the norm could be for instance the mean of the numbers of words of all the articles).

- The visible *reactivity degree*. It can be evaluated by the normalized number of visitor comments on the article (the norm could be for instance the mean of the numbers of visitor comments of all the articles).

- The *accessibility* of the news on the site. It is the location of the article on the Web page (top, bottom), the location of the Web page on the website (for instance the minimal number of links to click before reaching the article from the main page).

The value of an hyperlink cited by W_j in article A_r is in this case defined as $v_{r,j} \times Comp(\lambda_j, \mu_r)$.

3.2   Expressing the source network

No matter the choice of the method to evaluate the hyperlinks, the value $m_{i,j}$ in the weighted adjacency matrix M may be interpreted as the visible importance of the citation relation of the website W_i by W_j, and is computed by summing the values of the hyperlinks from W_j to W_i. In other words, if $E_{i,j}$ is the subset of the articles containing a link from W_i to W_j, according to the fourth evaluation method,

$$m_{i,j} = \sum_{A_r \in E_{i,j}} v_{r,j} Comp(\lambda_j, \mu_r)$$

4.0   BULDING A SOURCE NETWORK

The proposed model gives a major importance to the network of sources in information dynamic. In order to control precisely the way this network is built, we chose to develop a tool to extract data from the Web according to our needs. We named our tool ONICS (*Outils de Navigation, d'Indexation et de Classement des Sources*) which stands for browsing, indexing and sorting tool for sources.

4.1   Motivations

The objective of the extraction process is to be able to obtain a network of sources as close as possible to the reality of the field.

Real-world graphs result from concrete cases and are used in various fields [10]. Many come from computer science (the Internet network, the Web, peer-to-peer networks) but also from biology and human sciences.

These graphs do not come from an abstract representation but from the reality of the field. The source network we consider matches the definition of a real-world graph; one of its obvious characteristics is that it is not directly available. Indeed it is impossible to know every sources of information of each person publishing on the Web. Therefore we will only use a partial representation of such a network.

Hyperlinks on the Web provide directed relationships between two Web pages. Extracting hyperlinks leads to the establishment of a network between Web pages. Many studies about the Web are based on such a network. For example, the PageRank algorithm [12] uses it to compute authority scores from Web pages.

Classically, the extraction of Web resources is done with a Web crawler [3]. Web crawlers are usually recursive and differ in the process of hyperlinks selection and the choice of stopping conditions. Web crawling strategies have to be chosen carefully depending on the motivations and application field [2]. This section sums up the main choices of our own crawling process.

4.2   Extraction choices

4.2.1   Source centered extraction

We decided to center the extraction process on a list of interesting sources. This choice is consistent with intelligence analysts practices. Indeed, they monitor regularly a list of precise sources.

The crawling process aims to gather all the articles published by the selected sources. The crawling expansion is controlled by the user who decides which new sources to add. When a link is extracted from an article, the process does not automatically visit the cited link recursively. Thus, an extracted hyperlink will be processed only if the source citing this hyperlink presents an interest for the user and in that case the crawler will process every article from this source. This choice has the advantage to provide a consistent and complete source network (when a source is added, all its cited hyperlinks are extracted) and the network obtained does not depend on stopping conditions of the crawler.

From a practical point of view, the RSS feeds are used to obtain the list of published articles in a certain period of time. The feeds give the URL of Web pages containing articles and ensure that we retrieve all the articles published by a website on a given period.

4.2.2   Hyperlinks selection by articles selection

In order to validate our interpretation of a hyperlink as a *citation link*, we chose to select only the external hyperlinks cited in the body of the articles. Thus, we removed commercial links, those being a part of the internal browsing structure of the website as well as those in Web users comments.

A generic process to extract the text of an article from its Web page has been developed. It is an integral part of the extraction process and it uses the URL of a Web page containing an article as an input. It is based on the Document Object Model (DOM) of the HTML page and a series of heuristics. These heuristics use different thresholds defined manually. The following points describe the main steps of the process:

- It selects each text node with more than 100 letters. This choice is a simple way to keep the text of the page being a part of a phrase as it removes the usually small pieces of text from the menus, caption and commercials.

- Then it goes up in the DOM tree of 1, 2 and 3 levels to identify a list of parent nodes. These parent nodes are *de facto* an aggregation of the text nodes using the tree structure of the page. A going up of a maximum of 3 levels is a way to aggregate in most of the cases the totality of the article text without merging it with the readers comments at the same time.

- Among those parent nodes, it selects those containing text nodes of which the sum of letters is bigger than 500. This selection defines the minimal size an article must have and removes partly the users comments that may have been included.

- If the previous step returns several results, the process chooses among the results the node with the first position in the tree. Practically, in some cases, user comments may still be among the selected nodes. The position of the node in the page is a satisfactory criterion to differentiate the article from comments.

The text of the article is then the concatenation of the content of all the text nodes of the selected node. Eventually, the hyperlinks selection returns the set of all the links cited in the extracted text.

4.3   Extraction process

Figure 1 illustrates the overall extraction process. The data is stored in a database of three tables containing the sites to crawl, the extracted articles and extracted links. The first step of the process consists in loading all the websites from the table *Sites*. For each website, we have the URL of its RSS feed, its name and its URL. The second step connects to the Web and extracts from the RSS feed the list of recently published articles. It also gives the title of each article, its author and publication date. From the URL of the articles, the third step gets after a second connection the text of the article as described in the previous section and the hyperlinks cited in the text. Finally, the obtained data (links, article text and metadata) are saved in the database during the fourth step.
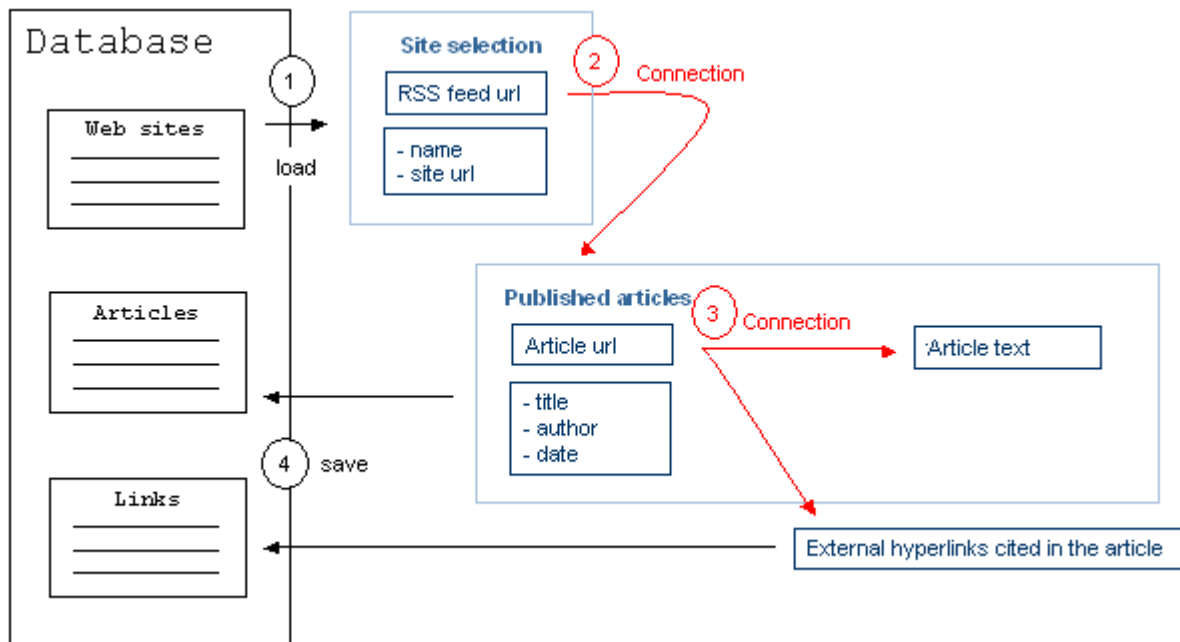

Figure 1: Extraction process

4.4   Extraction results

The extraction process collects citation links between websites but also the text of the article and some metadata including the publication date. Thus, some statistics can be computed from the database.

Currently, the database contains more than 60000 articles and 40000 links for a total of 110 websites crawled daily between February and July 2009. For this database was created for testing purposes, we chose to crawl a series of generalist websites. Figure 2 gives an insight of the database and some statistics that can be computed from it. The columns contain in order the name of the website, its number of published articles, the number of hyperlinks it has cited, the ratio between the number of links and the

number of articles, the average number of articles published daily, the daily average of link citations and the mean interval between publications.

It is worth noting that traditional news websites (like lefigaro, lemonde or lesechos) publish very frequently but use only a few hyperlinks. On the other hand, news websites coming from the Web (such as electronlibre, rue89 or techcrunch) publish less often but use numerous links.

n character var	nb_articles bigint	nb_links bigint	ratio numeric	avg_art_day numeric	avg_lin_day numeric	pub_mean interval
electronlibre	525	476	0.90	5.14	4.66	05:46:00
kassandre	56	69	1.23	1.19	1.46	4 days 10:58:52
lefigaro	1278	765	0.59	22.82	13.6	01:57:58
lemonde	1658	608	0.36	21.81	8.00	01:41:44
lesechos	1479	332	0.22	20.26	4.54	01:54:18
marianne2	34	47	1.38	1.88	2.61	1 day 19:02:12
rue89	484	1340	2.76	5.50	15.22	05:49:55
techcrunch	878	2690	3.06	13.93	42.69	02:04:54

Figure 2: Insight of the database and statistics computed

The publication date is very important because it will be used to build source networks on a given period of time. By varying the temporal parameter, we obtain a dynamic network of which the variations can be interpreted as a characterization of information dynamics.

4.5   Extraction evaluation

In order to have a first insight on the efficiency of our extraction tool, we established a simple evaluation of the quality of the data gathered by focusing on the central part of the tool, the article extraction process. We chose to analyse the extracted texts on two criteria: the presence of the text of the article (which can be complete, partial or not present) and the presence or not of additional and non-relevant text (noise). We chose 115 articles published by 99 different websites of our database and compared manually the article text we retrieved with the targeted article text of the Web page. Table 1 contains the number of articles we sorted in each of the 6 categories (identified as letters from A to F).

	Complete article	Partial article	Article not present
Noise	A=15	B=0	C=5
No noise	D=79	E=4	F=12

Table 1: Results of the evaluation of article extraction

First of all, the obvious success case is when the article is extracted completely and without noise. Therefore, we define the success rate of the extraction as the percentage of texts sorted in category D. This rate is around 70%. In our problem of hyperlink extraction, we may accept complete article texts with noise if we consider that practically, the additional text is usually some text "close to" (in the Web page display) the article and does not contain external hyperlinks. In that case (A+D), the success rate reaches 82%.

The failure rate defined with the number C+F is around 15% which seems quite high at first sight. Nevertheless, it has to be moderated because in most cases the failure in identifying the article happens when the article is very small and is easily explained by our choice to define a minimal article size of 500 letters. Further experiments could be carried out in order to check the effect of reducing this threshold, but at this point we consider that the relevancy of article that small is questionable and that the number of hyperlinks present in these articles is negligible.

This simple evaluation reveals that our article extraction process seems efficiently enough for our problem of link extraction.

5.0 ANOMALY DETECTION

Given the data gathered by our extraction process and in order to carry out a first series of experiments with extracted data as simple as possible we chose to represent the source network by an adjacency matrix built using the basic evaluation described in section 3.1. From this matrix we propose a method to detect anomalies in information diffusion. We call an *anomaly* a substantial variation (for example a consistent increase in the number of citations) between a source network serving the purpose of a reference and the source network we want to study. This work is a first step towards the analysis of information flows. Practically, such a tool points out to analysts unusual behaviours in information diffusion that may reveal informational phenomena such as amplification phenomena.

The process of anomaly detection consists in comparing a reference network with the one we want to analyse.

5.1 Reference network and analysed network

In this section we introduce $M_{ref} = \left(m_{i,j}^{ref} \right)$ the matrix representing the network used as a reference to analyse the network represented by the matrix noted $M = \left(m_{i,j} \right)$.

Practically, these two matrices are built from the data gathered during a given period of time. In order that the reference network be representative enough, the reference period duration has to be much longer than the one of the period to be analysed.

The matrix M^{ref} is then normalized by the ratio between the period analysed and the reference period. Our experiments have been carried out with an analysed period of one week and a reference period defined as the ten weeks before the analysed period.

5.2 Methodology

The main idea of the algorithm is to compare M and M^{ref} in order to point out values which changed too much and which may constitute anomalies.

First of all we need to find a way to compare M and M^{ref}, and more precisely a way to compare the current value $m_{i,j}$, to the reference value $m_{i,j}^{ref}$.

Rather than making a simple difference we strived to find a function φ that would represent a real relative change, having the following characteristics:

- φ has two variables, and somehow φ (x,y) represents the relative change of value x compared to the reference value y.

- φ(x,y)=0 if x=y.

- φ(x,y) is low if the relative difference between x and y is low, and important otherwise (e.g. φ(105,100)<< φ (10,5)).

- φ (x,y) grows with x.

For these reasons we define the function φ as:

$$\varphi(x,y) = \frac{|x-y|}{1+y}$$

The main two steps of the anomaly detection process are:

- An aggregation of the values in each matrix to obtain only one value for each website representing the way each website is cited globally. We used a sum to aggregate these values and we obtain two vectors D and D$^{\text{ref}}$ such as: $D = (d_i)_{i \in I} = \left(\sum_{j \in I} m_{i,j} \right)_{i \in I}$ and $D_{ref} = (d_i^{ref})_{i \in I} = \left(\sum_{j \in I} m_{i,j}^{ref} \right)_{i \in I}$.

- The calculation of the relative change by applying the φ function to D and D$^{\text{ref}}$. We obtain a vector D' such as $D' = (d'_i)_{i \in I} = \left(\varphi(d_i, d_i^{ref}) \right)_{i \in I}$. D' represents the total changes in the way each website is cited.

Finally, we focus on the set of sites W' for which this value changed the most:

$$W' = \left\{ W_i / d'_i \geq \bar{d} + \sigma(d) \right\}$$

where $\bar{d}$ is the average of every d'$_i$ and σ(d) its standard deviation. W' contains what we call the anomalies.

5.3.2   Results

We carried out our experiments on three networks extracted from data gathered over one-week periods. Table 2 contains the results of these experiments. The first line contains the date of the first day of the studied period; the date in brackets is the first day of the reference period that lasts 10 weeks.

A first result shows the website of the World Health Organization (WHO) for the weeks starting April 20th and April 27th. During this period, the WHO announced that the swine flu may turn into a pandemic. This first result reflects the media frenzy that followed this announcement. A request on Google trends [1] with the keyword *flu* returns a very specific peak on April 27th, which confirms the media frenzy on this subject on this period.

On the week of April 20th, the anomaly detection process detects the website of the World Digital Library (wdl) and the websites of Google news timeline and similar-images (newstimeline.googlelabs, similar-images.googlelabs). They were detected because those websites were respectively launched on April 21st and 20th.

The list of the websites detected during the week of May 18th contains the whitehouse website. It can be explained by the fact that Obama Administration launched several open government initiative websites at this time.

Dates	04/20/09 (02/08/09)	04/27/09 (02/15/09)	05/18/09 (03/08/09)
Web sites	computerworld ecrans liberation **newstimeline.googlelabs** **similar-images.googlelabs** team.lejdd theregister.co **wdl** **who** wired	bloomberg businessweek computerworld digitimes eff elmundo flickr france-info googlepublicpolicy.blogspot microsoft online.wsj reuters wired **who**	amazon elpais engadget facebook i.dailymail.co itunes.apple lexpansion maps.google midilibre pcworld telegraph.co thesun.co translate.google washingtonpost **whitehouse** wired youtube

Table 2: websites pointed out by the anomaly detection process for three different periods

Table 3 contains the results when we tried to apply the φ function before the aggregation of the values of the matrices.

Dates	04/20/09 (02/08/09)	04/27/09 (02/15/09)	05/18/09 (03/08/09)
Web sites	amazon computerworld dailymotion ecrans en.wikipedia facebook fr.wikipedia googlecom guardian.co lefigaro liberation **newstimeline.googlelabs** nytimes online.wsj **wdl** youtube	assemblee-nationale dailymotion eff en.wikipedia facebook flickr fr.wikipedia googlecom guardian.co lefigaro lemonde liberation microsoft news.cnet nytimes online.wsj **who** wired youtube	amazon en.wikipedia facebook fr.wikipedia googlecom guardian.co lefigaro news.bbc.co news.cnet nytimes online.wsj pcworld wired youtube

Table 3: websites pointed out when applying φ before the aggregation

It appears that some websites are detected almost all the time. Those sites are usually not very useful because most of them are not news websites (google, wikipedia, youtube, facebook, flickr) or are very generalist (liberation, lefigaro, nytimes). They are nevertheless very popular. This result may be explained by the use of a sum to obtain the global changes for each site. In the case of websites usually very cited, for every period, small variations in citations occur and the sum of these variations is regularly high enough to be detected as an anomaly.

These first results show that the choice of the order to apply φ and the sum is important and validate partly our anomaly detection process. Nevertheless, the method is probably perfectible as many website detected during the experiments are still sorely interpretable.

6.0 CONCLUSION

The Internet is a very large and complex information diffusion medium for open sources. Its very network structure is playing a major role in the dynamics of information propagation.

This study presents an approach to tackle the issue of anomalies detection of information diffusion based on a network of sources. It introduces different levels of link representation between sources so that the model can be adapted to the data available to the user and gives at the same time a framework that formalize information dynamics.

We introduced an extraction process that is able to gather the basic data needed by our system.

Such a system can be used as a decision support tool to alert intelligence analysts in the case of abnormal information diffusion behaviours. We obtained promising results by using a very simple link representation method. It is worth pointing that numerous changes on the system are possible: in the detection process, the φ function, as well as the aggregation operator and the threshold definition can be modified. Thus, the system is adaptable to any type of anomalies (such as amplification phenomena).

Moreover, it is possible to imagine a system that would identify patterns of anomalies and would help analysts to interpret changes in information diffusion.

REFERENCES

[1]   Google trends. available at http://www.google.com/trends.

[2]   C. C. Aggarwal. On learning strategies for topic specific web crawling. *Next Generation Data Mining Applications*, January 2004.

[3]   C. Castillo. Effective web crawling. *SIGIR Forum*, 39(1):55–56, 2005.

[4]   P. S. Dodds and D. J. Watts. Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92(21), May 2004.

[5]   P. S. Dodds and D. J. Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232(4):587–604, February 2005.

[6]   D. Gruhl, D. Liben-Nowell, R. Guha, and A. Tomkins. Information diffusion through blogspace. *SIGKDD Explor.Newsl.*, 6(2):43–52, December 2004.

[7]   A. Java, P. Kolari, T. Finin, and T. Oates. Modeling the spread of influence on the blogosphere. Technical report, University of Maryland, Baltimore County, March 2006.

[8]   D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146. ACM Press, 2003.

[9]   D. Kempe, J. Kleinberg, and E. Tardos. Influential nodes in a diffusion model for social networks. In *ICALP*, pages 1127–1138, 2005.

[10]   M. Latapy and C. Magnien. Measuring fundamental properties of real-world complex networks. In *CoRR*, Sep 2006.

[11]   I. Ounis, M. de Rijke, C. Macdonald, G. A. Mishne, and I. Soboroff. Overview of the trec-2006 blog track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*. Nist, 2007.

[12]   M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in PageRank. In *Advances in Neural Information Processing Systems 14 (NIPS)*, pages 1441–1447. MIT Press, 2001.

[13]   X. Song, Y. Chi, K. Hino, and B. Tseng. Identifying opinion leaders in the blogosphere. In *CIKM'07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 971–974, New York, NY, USA, 2007. ACM.

[14]   D. J. Watts, R. Muhamad, D. C. Medina, and P. S. Dodds. Multiscale, resurgent epidemics in a hierarchical metapopulation model. In *Proc Natl Acad Sci U S A*, 102(32):11157–11162, August 2005.