

System and method for detecting reputation attacks

FIELD

[0001] The technology relates to the field of computing, in particular to systems and methods for detecting reputation attacks.

[0002] A reputation attack within the context of this specification is a method of influencing public opinion, carried out by posting information in open Internet sources, in particular, texts discrediting the reputation of the object of attack. In other words, the purpose of the reputation attack is to form a negative attitude of the audience towards the object of attack, through placement of specific publications on the Internet.

[0003] An object of the reputation attack may be, as a non-limiting example, a person, i.e. specific person; organization; project, for example, such as construction of the Crimean bridge; brand such as Adidas or Pyaterochka; territory or country; an event or activity, such as the Scarlet Sails Graduation Party; technology or product, e.g. the Sputnik V vaccine or Angara space rocket.

BACKGROUND

[0004] Ways of influencing public opinion, in particular ways of worsening or improving someone's reputation, have been known to mankind since ancient times. However, advent and development of the global Internet as a means of mass communications has given rise to a whole layer of new methods and techniques for manipulating public opinion. Pursuing goals, that are ancient as human society itself, these manipulation methods, however, are often technically new. This, in turn, gives rise to the need for use of technically new means and methods for at least revealing such manipulations.

[0005] The prior art discloses the publication “REPULSE OF INFORMATION ATTACK: ALGORITHM OF ACTIONS” (D. Shubenok, I. Ashmanov, publ. May 28, 2018) posted at the time of this application filing at <https://www.ashmanov.com/education/articles/otrazhenie-informatsionnoy-ataki-algoritm-deystviy/>.

[0006] This publication is rather descriptive; it specifies which namely modern tools could be in principle used for reputation attacks but does not disclose specific approaches for detecting such attacks. Besides, this publication contains a specification of only one of the possible scenarios of reputation attack, while there are quite a few such scenarios, and detection of attacks that differ from the described scenario often requires taking into account the factors other than those specified by the authors.

[0007] Nevertheless, this publication discloses a fact that is quite important in the context of this application, namely, that a reputation attack in the overwhelming majority of cases is carried out not by

a single actor, but massively, using a significant number of accounts, often automatically controlled by special programs (bots).

[0008] The prior art also discloses the patent RU2656583C1, "SYSTEM OF COMPUTER AIDED ANALYSIS OF FACTS" (JSC "Kribrum", publ. 06/05/2018), disclosing a system for checking and analyzing the behavioral actions of users in social media. The technical result of the corresponding method consists in improving the efficiency of automated detection of behavioral risks of social media users.

[0009] In other words, although this system relates to the systems aimed at identifying ways of influencing public opinion, its main function is to analyze the publications of social network users and determine their skills, and also the level of threat that a specified user may constitute. Methods for detecting reputation attacks are not disclosed in this patent in contrast to the method described below.

[0010] Moreover, the prior art publication US20110113096A1, "System and method for monitoring activity of a specified user on Internet-based social networks" (Profile Protector LLC, publ. 05/12/2011) discloses a system and method for monitoring activity on a social network. The monitoring criteria are predefined by the client to monitor the activity on the page of a certain user on a social network. Activity monitoring access to the page of a specified user on a social network is established through a social network application programming interface based on predetermined identification information that identifies a specified user on a social network. The client is notified when the monitored activity meets at least one of the predefined monitoring criteria.

[0011] It is easy to see that this publication is also devoted to the analysis of activity of a predefined account (social network user) and does not disclose detecting the fact of reputation attack, in contrast to the method described below.

[0012] Based on the results of the prior art study, it may be concluded that there is a need for a technical solution eliminating the disadvantages of the above approaches. The solution described below is designed to solve at least some of the problems revealed in the prior art analysis.

SUMMARY

[0013] The object of the proposed technology consists in development of a method and system for detecting reputation attacks.

[0014] The technical result of the claimed technique is automated detection of the fact of reputation attack, and also timely notifying the responsible persons of attack detection.

[0015] This technical result is achieved due to the fact, that the method for detecting reputation attacks, executed by a computing device, comprises the following steps at preliminary stage: crawling the Internet and finding publication sources, identifying the sources used for reputation attacks among

the found publication sources, finding accounts from which entries have been posted in the identified publication sources, identifying among the found accounts the ones controlled by bots, storing the obtained data on sources used for reputation attacks and accounts controlled by bots in the database; then, at operational stage: obtaining words and phrases characterizing the object of the reputation attack, crawling the Internet and finding publications comprising words and phrases characterizing the object of the reputation attack, extracting hyperlinks from the found publications, computing quantitative characteristics of the publications and dynamics of their change using data on sources used for reputation attacks and accounts controlled by bots, computing the parameters characterizing the probability of reputation attack based on the computed quantitative characteristics, and in response to exceeding the preset threshold value by at least one computed parameter determining, based on the computed parameters, type of attack and level of attack, generating and sending a notification of the reputation attack, and also of the attack type and level.

[0016] This technical result is achieved due to the fact, that the system of detecting reputation attacks, configured to crawl the Internet, comprises at least a processor and also a storage device comprising at least one database, and also machine-readable instructions, which, if being executed by the processor, implement the described method.

[0017] In the specific embodiment the method is characterized in that at least the following sources relate to the publication sources used for reputation attacks:

- compromising material aggregators,
- social networks,
- data leak aggregators,
- advertising platforms,
- groups of related sources,
- user feedback aggregators,
- sites for hiring remote workers.

[0018] In the other specific embodiment the method is characterized in that the groups of related sources include the groups of sources that have posted identical publications at least a predefined number of times with a publication time difference not exceeding a predefined threshold value.

[0019] In the other possible embodiment, the method is characterized in that accounts controlled by bots include the accounts that have made at least a predefined number of publications within a predefined period of time.

[0020] In the other possible embodiment, the method is characterized in that the quantitative characteristics of publications include at least the following values:

- total number of publications,

- number of posts made by bots,
- number of publications made on compromising material aggregators,
- number of publications made by groups of related publication sources,
- number of publications made by groups of related sources which are also compromising material aggregators,
- number of publications made on advertising platforms,
- number of publications made on advertising platforms included in the group of related sources,
- number of publications made on user feedback aggregators,
- number of publications made on data leak aggregators,
- number of publications made on sites for hiring remote workers,
- total number of publications duplicating each other,
- total number of publications on compromising material aggregators duplicating each other,
- total number of publications on compromising material aggregators duplicating each other and made by bots,
- total number of links duplicating each other,
- number of accounts from which the found publications were posted,
- number of accounts, controlled by bots, from which the found publications were posted,
- number of accounts from which the publications found on compromising material aggregators were posted,
- number of accounts, controlled by bots, from which the publications were posted on compromising material aggregators,
- number of accounts from which the publications found on advertising platforms were posted.

[0021] In the other possible embodiment, the method is characterized in that the quantitative characteristic change dynamics is computed based on the value of these characteristics computed within the preset time interval with the preset increments.

[0022] In the other possible embodiment, the method is characterized in that the parameters characterizing a probability of reputation attack are computed for each quantitative characteristic as absolute, expressed in units, and relative, expressed as a percentage, difference between adjacent values of this characteristic.

[0023] In the other possible embodiment, the method is characterized in that the notification of reputation attack could have a numerical expression characterizing the level of attack intensity.

[0024] In the other possible embodiment, the method is characterized in that the notification of reputation attack could have one of three levels: "Warning", "Threat", "Attack".

[0025] In the other possible embodiment, the method is characterized in that at least one generated notification of reputation attack is transmitted by at least one of the following communication methods:

- e-mail,
- SMS,
- MMS,
- push notifications,
- instant messenger messages,
- API events

[0026] The technical result of the claimed technique may also be achieved by a method for detecting reputation attacks, executed by a computing device, having steps: obtaining words and phrases characterizing an object of a reputation attack, crawling the Internet and finding publications comprising words and phrases characterizing the object of the reputation attack, extracting hyperlinks from the found publications, computing quantitative characteristics of the publications and dynamics of their change using data on publication sources used for reputation attacks and data on accounts from which entries are posted in the publication sources used for reputation attacks, the accounts being controlled by bots, computing parameters characterizing a probability of the reputation attack based on the computed quantitative characteristics, and in response to exceeding a preset threshold value by at least one computed parameter, determining, based on the computed parameters, a type of the attack and a level of the attack, and generating and sending a notification of the reputation attack, and also of the attack type and level.

BRIEF DESCRIPTION OF THE DRAWINGS

[0027] The accompanying drawings, which are included for additional understanding of the technology and form a part of this specification, illustrate the technology embodiments and together with the specification serve to explain the principles of the technology.

[0028] The claimed technology is illustrated by the following drawings, wherein:

[0029] Fig. 1A illustrates a flow chart of the preliminary stage of the described method.

[0030] Fig. 1B illustrates a flowchart of one of the preliminary stage steps of the described method.

[0031] Fig. 1C illustrates a flowchart of the other preliminary stage step of the described method.

[0032] Fig. 1D illustrates a flowchart of the other preliminary stage step of the described method.

[0033] Fig. 2A illustrates a flow chart of the operational stage of the described method.

[0034] Fig. 2B illustrates a flowchart of one of the operational stage steps of the described method.

[0035] Fig. 2C illustrates a flowchart of the other operational stage step of the described method.

[0036] Fig. 3 illustrates a block diagram of one of the possible algorithms for evaluations of the attack method and the attack nature.

[0037] Fig. 4 illustrates a non-limiting example of the computing device schematic diagram.

DETAILED DESCRIPTION

[0038] Description of the exemplary embodiments of the claimed technology is given below.

[0039] The subject matters and features of the present technology, methods for achieving these subject matters and features may become apparent by reference to exemplary embodiments. However, the present technology is not limited to the exemplary embodiments disclosed below, and it may be implemented in various forms. The essence given in the specification is nothing, but specific details provided to assist a person skilled in the art in a comprehensive understanding of the technology, and this technology is defined only within the scope of the appended claims.

[0040] In the subsequent specification of the method and system for detecting reputation attacks, the following basic terms and definitions are used:

[0041] Account – is a unique account, which creation is a necessary and sufficient condition for a specific user to participate in communications through this website or this social network. It is characterized by availability of a user identifier, unique within the framework of this website or social network: a username, its sequence number or other combination of characters.

[0042] Social network – is an Internet platform that allows the registered (having an account of this network) users to communicate with each other. The content on such a site is created by the users themselves. In terms of the user interface, a social network could be as a website, such as vk.com, facebook.com, as an instant messenger, Internet messenger, such as Telegram or Discord.

[0043] Source or source of publications (in this case) – is a website or community (channel, group, server) on a social network that specializes in posting texts. Within the scope of this specification, the sources include:

- Mass media, on which websites there could be both publications as such and comments under the publications;
- forums;
- blogs of journalists, politicians and public figures;
- communities (groups, publics) on social networks;
- video hosting services and stream servers;

- question and answer services;
- sign-in services;
- crowdfunding services;
- websites performing the following functions:
 - a. user feedback aggregators,
 - b. rating agencies;
 - c. "bulleting boards", including:
 - i. account exchanges,
 - ii. sites for hiring remote workers.

[0044] "Bulleting board" (in this case) – is a website providing services for posting advertisements on a specific or arbitrary topic.

[0045] Account exchange (in this case) – is a kind of "bulletin board" where offers for sale or lease of accounts owned by people or bots, and also messages about desire to purchase or lease such accounts, are posted.

[0046] Bot (in this case) – is an account controlled by a program that is configured to leave messages on behalf of one of the users of a specified social network. Usually, after the initial setup, the bot acts autonomously and posts specific content messages on a specified social network without operator's participation.

[0047] A group of related sources (in this case) – is a group of sources, where texts are posted by one person or one organized group of people.

[0048] Compromising material aggregator (in this case) – is a source that posts only the texts of compromising material nature. An example of such a source is compromat.ru website.

[0049] Data leaks aggregator (in this case) – is a source that posts only the texts of data leak (insider) nature. An example of such a source is WikiLeaks website.

[0050] Rating agency (in this case) – is a website, which main functionality is to form and display the rating of certain specialization websites. For example, rating of the most influential user feedback aggregators, rating of account exchanges, rating of SMM service exchanges, etc.

[0051] Advertising platform (in this case) – is a source that is a mass media outlet that posts news but differs in that it allows placement of arbitrary content text under the guise of another piece of news as an advertisement.

[0052] It is also worth noting that in the context of the present specification, unless clearly specified otherwise, the words the "first" and the "second" are used solely to distinguish the nouns to which they refer to from each other, and not for the purposes of describing any specific relationship between these nouns.

[0053] First, preliminary stage (100) is performed to implement the described method for detecting reputation attacks, as described below with reference to Fig. 1A.

[0054] Preliminary stage (100) begins with the step (110), which includes the Internet crawling and finding web pages containing publications. Crawling is performed in any well-known way, using any program implementing the functions of a web parser, i.e. an automatic “collector” of publications from various websites, such as CloudScrape or Scrapinghub. In one embodiment of the described method, prior to crawling, language or languages in which the publications are to be written (for example, Russian, or Russian and English) are set. In the other possible embodiment, the search is carried out without language limitation.

[0055] It is also possible such embodiment of the described method, wherein the publication sources are additionally retrieved from the web pages obtained during the aforementioned collection of publications from various websites. Automated processing (parsing) of such web pages, for example, performed by the preliminarily prepared script, could be used to extract links to publication sources from them and replenish the general list of publications with the extracted links.

[0056] It is also possible such embodiment of the method, wherein the publication sources are additionally found by analyzing e-mails, including unsolicited emailing (spam). This can be done by any well-known method. For example, a number of e-mail accounts may have been registered in advance, the addresses of which may have been put out in the open. Such addresses, as a rule, soon make the mailing lists (spam), and these addresses begin to receive e-mails, including those containing links to various publication sources listed above. The automated processing (parsing) of such e-mails, for example, performed by the preliminarily prepared script, could be used to extract links to publication sources from them and replenish the general list of publications with extracted links.

[0057] Step (110) results in a list of detected web pages stored in the database.

[0058] At this point, step (110) ends and the method proceeds to step (120), where the found web pages are analyzed, with selecting and storing at least: title, (author's) account, hyperlink (URL) to the web page, time of being put out in the open (time and date of publication), its source, for example, formed by truncating a hyperlink to the second or third level domain name, as well as the publication text as such. Such identification of the listed data types on a web page is one of the typical functions of web parsers and it could be performed using the program used. Alternatively, the selection of the named fields could be performed by the preliminarily prepared script implementing any well-known algorithm.

[0059] For example, as a result of the above step (120) a publication titled “Attention!” comprising the text: “I heard there will be imposed a pet tax soon!”, published from the sampleuser account, with the date and time of publication as 02/11/2021 17:21:35, a hyperlink to this publication: <http://www.livejournal.com/sampleuser/12345678.html> and also the source: sampleuser.livejournal.com could be stored in the database.

[0060] Then the method proceeds to step (130), where at least the following types of sources are identified among the found publication sources: social networks, compromising material aggregators, data leak aggregators, sign-in platforms.

[0061] It is worth noting that constant use of one and the same domain name is typical of all the listed sources. Generally, a significant portion of such sources budget is advertising revenue; quite often they run their own advertising campaigns attracting new users. Therefore, the domain names of such sources remain ones and the same for years, which, in turn, enables to have permanent lists of domain names and to check membership of the next source in one of the named types according to them.

[0062] For example, there could be separate lists, for example, "Social Networks" list, which stores such domain names as facebook.com, vk.com, livejournal.com, etc., "Compromising Material Aggregators" list, which stores domains names like compromat.ru or compromat.livejournal.com, "Data Leak Aggregators" list, which stores domain names like wikileaks.com, and also "Sign-in Platforms" list, which comprises domain names like change.org, democrator.ru, e-petition.am, etc.

[0063] At step (130), each next found publication source is checked in turn for presence in each of the specified lists. In case of a match, the source being checked is appropriately classified, that is, a tag, corresponding to the list where its domain name was found, is assigned in the database for this source.

[0064] Thus, in the above example, for a publication found at <http://www.livejournal.com/sampleuser/12345678.html>, and also for all the other publications found on the livejournal.com domain, the tag "Social Networks" will be assigned in the database, as livejournal.com domain name will be found in the "Social Networks" list.

[0065] It is worth noting that one and the same domain name or similar domain names could be in different lists. For example, the "Social Networks" list may contain livejournal.com domain name, while the "Compromising Material Aggregators" list may contain such domain names as slivaem-kompromat.livejournal.com, compromat.livejournal.com, etc. At the end of step (130), at least some of the publications and their corresponding sources, which will be found in the listed lists, will be classified. Technically, classification could be, for example, tags assigned in the database, with each of them corresponding to one of the source types: "Social Network", "Compromising Material Aggregator", etc. As already noted, the publication source may simultaneously refer to different source types, therefore, as a result of step (130) more than one tag could be assigned to the source.

[0066] Then, the method proceeds to step (140), where the groups of related publication sources are identified among the found publication sources.

[0067] All sources classified at step (130) are not excluded from further processing during the next step (140), since, for example, a social network group (public) could function as, for example, an advertising platform or be a part of a group of the related sources.

[0068] Hereafter, with reference to Fig. 1B, step (140) is described, where the groups of related publication sources are identified among the found publication sources.

[0069] Step (140) begins with selection (141) of the next publication from the found publications. Then, at step (142), it is checked whether there are duplicates of the selected publication among all found publications. In this case, a duplicate means a strict coincidence of the selected publication text with the text of any other publication.

[0070] Technically, step (142) consists in searching the database for all publications from the other sources, for which the text in the database “Publication Text” field is an exact copy of the text available in this field for the selected publication. Such a search could be carried out by in any well-known method, selected depending on the architecture of the database used.

[0071] If no duplicates are found at step (142), that is, there is no publication, which text matches the text of the selected publication, the method returns to step (141), where the next publication is selected.

[0072] If duplicates are found at step (142) , that is, at least one publication is found, which text exactly matches the text of the selected publication, the method proceeds to step (143).

[0073] At step (143), the group of duplicate publication sources, found at step (142), is assigned to the group of candidate sources. At the same time, the sources, which the found publications belong to, are stored as a separate list, and it is checked whether the publication time in all found publications is the same.

[0074] Time coincidence in this case may be fuzzy, when the publication time is considered to be coincident, if it differs in any direction from the time of the publication selected at step (141) by no more than a preset dT value, for example, by no more than 30 seconds.

[0075] The sources posted duplicate publications with the time difference dT greater than a preset value are excluded from the group of candidate sources.

[0076] For example, if at step (142) the following sources of duplicate publications, that posted publications at the specified time have been found:

- | | |
|------------------------------|---------------------|
| • sampleuser.livejournal.com | 11.02.2021 17:21:35 |
| • website.com | 11.02.2021 17:21:07 |
| • sample.newspaper.ru | 11.02.2021 17:21:59 |
| • examplechange.org | 15.02.2021 07:01:06 |

then as a result of the step (143) with preset dT value equal to 30 seconds, the following sources will remain in the list of the candidate source group:

- sampleuser.livejournal.com
- website.com
- sample.newspaper.ru

[0077] On completing the step (143), the method proceeds to step (144), where it is checked whether the found group is zero (empty). If a group of candidate sources is proven to be zero, that is, if all found publications are made by candidate sources with the time difference greater than dT , then the group of candidate sources is deleted, and the method returns to step (141), where the next publication is selected.

[0078] If the group of candidate sources is proven to be nonzero, that is, at least two candidate sources are found that have posted publications with the same text with the time difference not exceeding dT , then the list of the candidate source group is saved, the initial value $J = 1$ is assigned to the enumerative variable J , which value is stored associated with each such list, and the method proceeds to step (145).

[0079] At step (145) it is checked whether at least one of the candidate sources found at the previous step has been found again. In other words, it is checked whether at least one of the candidate sources found at step (144) is included in at least one list of candidate source groups previously stored. This is done by any well-known method, by sequential searching for each of the candidate sources found at step (144) in all the lists of candidate source groups previously stored.

[0080] If all candidate sources found at step (144) are absent in all previously stored lists of candidate source groups, that is, a group of candidate sources found at step (144) is new, the method returns to step (141), where the next publication is selected.

[0081] If at step (145) at least one candidate source included in at least one previously stored list of candidate source groups is found, the method proceeds to step (146).

[0082] At step (146), the lists of candidate source groups, where the same candidate sources have been found, are combined. For this purpose the following actions are performed: adding all candidate sources available in each list to the new combined list, and if a candidate source is in more than one list, it is not added again; then, storing the resulting combined list. Then all the values of the enumerative variable J associated with each of the found lists are summed up, and the resulting value J is assigned to the combined list. Then, initial lists are deleted, leaving the resulting combined list only.

[0083] For example, if during execution of step (145) in relation to the previously shown group of candidate sources, which had the value $J = 1$:

- sampleuser.livejournal.com
- website.com
- sample.newspaper.ru

one of these sources is found in the other previously stored list with the value $J = 3$, for example,

- anotherwebsite.es
- sampleuser.livejournal.com
- justasite.co.il

then, at step (146) these two lists will be combined into one list as follows:

- sampleuser.livejournal.com
- website.com
- sample.newspaper.ru
- anotherwebsite.es
- justasite.co.il.

and the enumerative variable J value, which is stored associated with this combined list, will be computed as the sum:

$$J = 1 + 3 = 4.$$

[0084] In that special case, if at step (145) a previously saved list, consisting of the same sources as the list created at step (144), is found, i.e, two completely identical lists are found, then the values of the enumerative variable J associated with each of the lists are summed, one of the lists is removed, and the resulting value J is assigned to the remaining list.

[0085] Then, the method proceeds to step (147), where the enumerative variable J value obtained at step (146) is compared with a preset threshold value Jmax. This preset threshold value is selected at the stage of setting up the system that implements the method. It has the meaning of the number of “group” publications posted at different times by overlapping or matching source groups, and it could be, for example, equal to 3.

[0086] If the value of the enumerative variable J is proven to be less than or equal to the set threshold value Jmax, the method returns to step (141), where the next publication is selected.

[0087] If the value of the enumerative variable J is proven to be greater than the threshold value Jmax, the method proceeds to step (148), where all publication sources included in the combined list are assigned to the group of related publication sources. In other words, the list for which $J > J_{max}$ is considered to be the list comprising a group of related publication sources. For all the publication sources included in it, the corresponding tag, “Group of Related Sources”, is assigned in the database, then, the method returns to step (141), where the next publication is selected.

[0088] The list for which $J > J_{max}$ is not deleted, it is still processed at step (140) along with all other lists of candidate sources, as described above.

[0089] Step (140) is executed cyclically until the end of the publication list, from which publications are selected at step (141). This completes the execution of step (140), and the method proceeds to step (150), as described above with reference to Fig. 1A.

[0090] At step (150), as it will be described later with reference to Fig. 1B, at least the following types of sources are identified in the found publication sources: advertising platforms, user feedback aggregators, account exchanges, SMM service exchanges and sites for hiring remote workers (freelance exchanges).

[0091] Step (150) begins with Internet search performed at step (151), during which the websites functioning as rating agencies are found.

[0092] This search is performed by well-known method, with the use of any well-known search engine such as Google. Preliminarily prepared sets of strings are used as keywords, enabling to generate a corresponding search query, for example, such as:

"rating of SMM exchanges"

"rating of account exchanges"

"rating of freelance exchanges"

"rating of the best user feedback aggregators"

[0093] Then, by analyzing the search results, which can be done by any well-known method, for example, using preliminarily prepared script, hyperlinks (URLs) to websites functioning as rating agencies are extracted, and storing these links in the form of lists, for example, a list of ratings of SMM exchanges, a list of ratings of account exchanges, etc. Thus, as a result of step (151), the lists of links to websites of rating agencies, sorted by the website activity specifics, are obtained.

[0094] At this point, step (151) ends, and the method proceeds to step (152), where the found sites of rating agencies are crawled. For this purpose, the lists of URLs generated in step (151) are used. Crawling is performed in any well-known way, using any program implementing the functions of a web parser, i.e. an automatic "collector" of publications from various websites, such as CloudScape or Scrapinghub.

[0095] As a result of step (152), the web pages of the crawled sites are obtained and stored in the database, which comprise, but not limited to, the ratings as such, i.e. ordered lists of websites functioning as account exchanges, SMM service exchanges, freelance exchanges and user feedback aggregators.

[0096] At this point, step (152) ends, and the method proceeds to step (153), where the lists of websites functioning as various exchanges, and also as user feedback aggregators, are generated. This is done by any well-known method that enables to extract links (URLs) to the sites listed in the ratings from the web pages of rating agencies stored at the previous step. The extracted links are stored in the lists, thus generating:

- the list of links to account exchanges,
- the list of links to SMM service exchanges,
- the list of links to freelance exchanges,
- the list of links to user feedback aggregators (1)

[0097] The lists generated in this way are stored, and at this point the method proceeds to step (154), where the lists generated at the previous step are analyzed and cleared. For this purpose, duplicate

entries are removed from the above lists, that is, duplicate links (URL) are excluded. Besides, at this step, the obtained links are truncated to the second level domain, so that the URL

`https://example-otzovik.su/index.html`

is found to be converted to the string

`example-otzovik.su`.

[0098] This can be done by any well-known method. As a result of step (154), four source lists corresponding to the list (1) are obtained and stored in the database.

[0099] Then, the method proceeds to step (155), where the found account exchanges are crawled. For this purpose, the list of URLs to the account exchanges, compiled at step (154), is used. Crawling is performed in any well-known way, using any program implementing the functions of a web parser, i.e. an automatic “collector” of publications from various websites, such as CloudScape or Scrapinghub.

[0100] As a result of step (155), the web pages of crawled sites are obtained and stored in the database, which comprise, but not limited to, the lists of accounts offered for sale or lease.

[0101] It is worth noting that accounts sold or offered for lease on account exchanges are intentionally controlled by bots. Therefore, at the next step (156), the web pages stored at step (155) are analyzed and names of the accounts offered for sale or lease are extracted from them, from which a list of accounts controlled by bots is generated. The analysis of web pages as such could be performed by any well-known method, for example, with the use of a script parsing a web page, extracting account names from it and storing them into a separate list.

[0102] The list of accounts controlled by bots before completing step (156) is used to tag the known accounts obtained at step (120) as “Bot”. Tags are assigned in the used database by any well-known method, in accordance with the database architecture used.

[0103] Those accounts that are present in the list obtained at step (156), but are absent in the database (i.e. accounts that are controlled by bots, but have not yet been found or have not been used) are also stored in the database with “Bot” tag and further are used along with all other known accounts.

[0104] An alternative embodiment of the described method is also possible, wherein step (156) is skipped, proceeding from step (155) to step (157).

[0105] The method proceeds to step (157), where the found freelance exchanges are crawled. For this purpose, the list of URLs of freelance exchanges, compiled in step (154), is used. Crawling is performed in any well-known way, using any program implementing the functions of a web parser, i.e. an automatic “collector” of publications from various websites, such as CloudScape or Scrapinghub.

[0106] As a result of step (157), the web pages of crawled freelance exchanges are obtained and stored in the database, which comprise, but not limited to, texts of tasks for freelancers to post reviews with the predefined focus on the pages of any given web resources.

[0107] It is worth noting that web resources on which freelancers are invited to post reviews with the predefined focus, as a rule, relate to the category of advertising platforms, i.e, they are online media that publish, in addition to regular news, paid publications with the predefined focus.

[0108] Therefore, at the next step (158), the web pages stored at step (157) are analyzed and links (URLs) to advertising platforms are extracted from them, from which a list of advertising platforms is generated. The analysis of web pages as such could be performed by any well-known method, for example, with the use of a script parsing a web page, extracting URLs from it and storing them into a separate list.

[0109] Then, the method proceeds to step (159), where the list thus generated is analyzed and cleared. For this purpose, duplicate entries are removed from the list, that is, duplicate links (URL) are excluded. Besides, at this step, the obtained links are truncated to the second level domain, so that the URL

`https://reklamnoe-smi.ru/index.html`

is found to be converted to the string

`reklamnoe-smi.ru.`

[0110] That is the end of step (150). The result of steps (130), (140) and (150) is classification of the data stored after step (110), that is, assignment of at least part of the found publication sources to at least one type of sources. As already noted, a publication source may simultaneously relate to different types of sources, more than one tag could be assigned to the source.

[0111] In one possible embodiment of the method, only the sources, which type was determined at steps (130), (140) and (150), are used for further data processing. In the other possible embodiment, all sources are used for further data processing.

[0112] Then, the method proceeds to step (160), where, in addition to the list obtained at step (156), the accounts controlled by bots are identified among the accounts found at step (120). Execution of step (160) will be described in detail below with reference to Fig. 1D.

[0113] Execution of step (160) as shown in Fig. 1D, begins at step (161), where publications posted on social networks are selected among the found publications. Since earlier, as a result of step (130), the publication sources, which are social networks, have been classified, execution of step (161) is a selection from the database of publications, which are marked as "Social network" in the database. Technically, such selection could be carried out by any well-known method selected depending on the architecture of the database used, for example, by sending a corresponding SQL query and receiving a response to it.

[0114] It is worth noting that "publications on social networks" in the context of this step (161) execution are the results of such actions, the implementation of which requires the most time

expenditures from the social network user. These include the following publications typical of modern social networks:

- entry (original message),
- comment (reply to someone's entry or comment),
- repost, that is, posting on its own behalf any entry made by an arbitrary user, indicating the sender's account and link to the original entry.

[0115] Such alternative means of social network users' expression, such as emoticons and likes/dislikes (votes "for" and "against") are not taken into account during this step.

[0116] After receiving publications on social networks, the method proceeds to step (162), wherein all publications made by one account are selected from the total array of publications. Since earlier, in the course of step (120), the accounts from which all publications were made have been identified, technically, step (162) is filtering the resulting array of publications by author (account). It can be executed by any well-known method, selected depending on the architecture of the database used. The account name as such is taken in turn from the general list of accounts stored in the database, and generated, as described earlier, at steps (120) and (156).

[0117] Prior to filtering the resulting array of publications (in Fig. 1D, this step is not shown for the sake of simplicity), there could be additional checking whether the given account has already the "Bot" tag. Such a tag could have been assigned earlier, during execution of step (156). If such a tag is present, step (163) is not executed, and it proceeds to the next account from the list of accounts.

[0118] Then, at step (163), the number M of publications made by this account at a given interval, for example, at an interval of 1 second or less, is computed. For this purpose, the publications are ordered by publication date and time using any well-known method, and then, the time intervals between each two publications adjacent in time are computed. For example, if an account has made publications conditionally designated as P1, P2, P3 and P4, then, the intervals between P1 and P2 publications, between P2 and P3, and also between P3 and P4 will be computed. Then, the number M of intervals, which duration is less than or equal to a preselected value, for example, less than or equal to 1 second, is computed.

[0119] Then, the method proceeds to step (164), where the computed M value is compared with a preset threshold. This threshold could be selected empirically at the stage of the system setup, and could be, for example, 4. If the M value for the analyzed account exceeds this preset threshold, the method proceeds to step (167), wherein this account is assigned to those accounts that are controlled by bots, and then returns to step (162).

[0120] If at step (164) the M value for the analyzed account is less than a preset threshold, the method proceeds to step (165), the period of time T within which the account made publications with at

least the set frequency F is computed. In this case, the F value could be selected in advance, at the stage of the system setup. For example, F could be selected as equal to one publication per hour or one publication per two hours.

[0121] For example, if an account has made publications conditionally designated as $P_1, P_2 \dots P_{400}$, then, at step (164) they will be ordered by the publication date and time, after that the time intervals between each two publications adjacent in time will be computed: between P_1 and P_2 , between P_2 and P_3 , and so on, up to the interval between P_{399} and P_{400} . Then, the time periods T_1, T_2, T_3 , etc., such that within each such period the frequency of publications exceeds the preset F value is found. In other words, all the time periods, during which the given account has posted publications more often than with the set frequency F , are found. This can be done by any well-known method.

[0122] Then, the duration of the time period T is determined as the maximum period duration among all the found time periods T_1, T_2, T_3 , etc.

[0123] Then, the method proceeds to step (166), where it is determined whether the duration of the time period T exceeds the preset threshold. For example, this threshold could be selected as equal to 36 hours or 48 hours. In other words, at this stage it is checked how long any publications have been posted from this account continuously, without a pause required for a person to sleep.

[0124] If T exceeds the preset threshold, the method proceeds to step (167), where this account is assigned to the accounts controlled by bots, that is, the tag “Bot” is assigned to it in the database, and then returns to step (162). Otherwise, if T does not exceed a preset threshold, the method returns to step (162).

[0125] For clarity of the block diagram in FIG. 1D, check of the condition “is the end of the account list reached?” is not shown, which could be performed each time before step (162), where the next account is selected for analysis. When this condition is met, this completes the execution of step (160), and the method returns to step (170), as described above with reference to Fig. 1A.

[0126] It is worth noting that steps (110), (120), (130), (140), (150) and (160) are shown in relation to the preliminary stage as a simple sequence for the sake of description simplicity. However, it is a possible embodiment of the system wherein these steps are executed not once, but cyclically, including execution in parallel with the steps that will be described below in relation to the operational stage. This activity could be carried out continuously, which enables to seed the databases constantly, and to have “fresh”, actual information in the database at any time.

[0127] At the final step (170) of the preliminary stage (100), all the information obtained at previous steps is stored in the database. This can be done by any well-known method. This completes the preliminary stage (100).

[0128] In order to implement the described method for detecting reputation attacks, upon completion of the preliminary stage (100), the operational stage (200) is executed as described below with reference to Fig. 2A.

[0129] The operational stage (200) begins with step (210), where at least one word or phrase, characterizing the object of the reputation attack, is obtained. This can be done by any well-known method. For example, a string preliminarily prepared in accordance with the format accepted by the system, comprising words and phrases characterizing the object of the reputation attack, could enter the method implementing system from the database, where tasks for the method implementing system are stored, after the system completes the previous task.

[0130] Also, any alternative embodiments of this step are possible without limitation, including import of words and phrases from the text of an email sent to a prearranged email address associated with the system implementing the described method, etc.

[0131] Then, the method proceeds to step (220). Before starting step (220), the system clock readings are obtained and stored in the database by any well-known method, i. e. current time at the step starting. Then, the Internet is crawled and web pages comprising the obtained words and phrases are found. Technically, this can be done by any well-known method, for example, in the same way as described above for step (110). Then, the method proceeds to step (230).

[0132] At step (230), the found web pages are analyzed and at least the following is selected: title, author's account, date and time, publication source, publication text; this is done in the same way as described above for step (120). The extracted information is stored in the database.

[0133] This completes step (230), and the method proceeds to step (240), where links (URL) are extracted from the found publication texts and lists of links are generated. This is done by any well-known method that enables to extract all links (URLs) from the publications stored at the previous step, for example, using preliminarily prepared script that finds such character combinations in each publication as http, https and www, and extracts the entire line starting with of these characters and ending with a space or "carriage return" or "line break" characters.

[0134] The links thus extracted are stored, for example, in the database, and the method proceeds to step (250). At step (250), the publications found at step (220) and links extracted at step (240) are analyzed, and the values of quantitative characteristics and their dynamics are computed.

[0135] The quantitative characteristics include at least the following: (2)

- total number of publications N ,
- number of publications made by bots, N_b ,
- number of publications made on compromising material aggregators, N_k ,
- number of publications made by groups of related publication sources, N_g

- number of publications made by groups of related sources which are also compromising material aggregators, Ngk
- number of publications made on advertising platforms, Nr,
- number of publications made on advertising platforms included in the group of related sources, Ngr,
- number of publications made on user feedback aggregators, No,
- number of publications made on data leak aggregators, Nu,
- number of publications made on sites for hiring remote workers, Nh,
- total number of publications duplicating each other, Nd,
- total number of publications on compromising material aggregators duplicating each other, Ndk,
- total number of publications on compromising material aggregators duplicating each other and made by bots, Ndbk,
- total number of links duplicating each other, Nld
- number of accounts from which the found publications were posted, Na
- number of accounts controlled by bots from which the found publications were posted, Nab
- number of accounts from which the publications found on compromising material aggregators were posted, Nak
- number of accounts, controlled by bots, from which the publications were posted on compromising material aggregators, Nabk
- number of accounts from which the publications found on advertising platforms were posted, Nar.

[0136] In this case, the dynamics of the said values change means the values computed within the preset time interval t with the preset increments (time interval between iterations) t_s . By way of non-limiting example, the interval t could be set as equal to 10 minutes, and the step t_s as equal to 1 minute.

[0137] Specific methods for computing the said values will be described in detail below, with references to Fig. 2B, Fig. 2C.

[0138] It is easy to see that the above quantitative characteristics could be conventionally combined into three main groups: characteristics that have the meaning of the number of certain publications, characteristics that have the meaning of the number of duplicates (repetitions) and characteristics that have the meaning of the number of accounts.

[0139] The characteristics having the meaning of the number of publications are computed as shown in Fig. 2B. Since during the preliminary stage, namely, at steps (130), (140), (150), (160), various sources of publications and accounts have been tagged, that is, at least some of them have been tagged in the database as “Compromising Material Aggregator”, “Group of Related Sources”, “Bot” and so on,

extracting the number of publications related to certain sources from the database is technically implemented as a search for entries with a corresponding tag in the database.

[0140] At the first step (251), at least one filtration criterion (tag) is selected. It is selected from the preliminarily prepared list of tags, for example, alternately selecting one tag after another.

[0141] Then the method proceeds to step (252), where a query to the database comprising the selected tag is made, and a list of publications corresponding to this query is obtained from the database, and at step (253) an estimate of the length of this list, i.e. the number of publications, is obtained. Then, at step (254), the obtained estimate as such is stored; and the obtained list itself could also be optionally stored.

[0142] As an example, there is a detailed description of computing the number of publications N_b made by bots. During the preliminary stage, namely, steps (156) and (160), the accounts controlled by bots have been identified, and each of these accounts has been tagged as “Bot” in the database.

[0143] At step (251), the “Bot” tag is obtained from the tag list; then, at step (252), a query is made to the database, and a list of publications found at step (220) and made from accounts tagged as “Bot”, is obtained from the database. This can be done by any well-known method, depending on the architecture of the database used, for example, by sending an appropriate SQL query.

[0144] Then, the length of the list is determined, i.e, the number of publications thus obtained. It will be the number of publications N_b made by bots. It is stored in the database; also, such list of publications could be optionally stored.

[0145] In the other example, in order to calculate the total number of found publications N at step (252), a query to the database may not be made, and N may be taken as equal to the total number of web pages stored at the current iteration of step (220). For example, at the first iteration of the method, at step (220), 100 web pages could be found, and N value = 100 will be stored in the database. At the second iteration of the method, the number of the found pages could become equal to 110, and N value = 110 will be stored in the database. At the third iteration of the method, the number of the found pages could become equal, for example, to 130, and N value = 130 will be stored in the database.

[0146] It is worth noting that the values of all quantitative characteristics computed at step (250) are stored in the database in the form of a vector, i.e. a sequence of numbers. For example, as a result of the above iterations, the following sequence of values will be saved for N characteristic:

$$N = (100, 110, 130).$$

[0147] In the other example, two tags are used to determine the number of publications made by groups of related sources, which are also the compromising material aggregators (Ngk): “Compromising Material Aggregators” and “Group of Related Sources”. When making a query to the database, these tags are combined by logical AND, thus obtaining a list of publications, where both of these tags have been assigned to each of the sources earlier, at steps (130) and (140).

[0148] Then, the length of the list is determined, i.e, the number of publications thus obtained. It will be the number of Ngk publications made by groups of related sources, which are also the compromising material aggregators. It is stored in the database; the list of publications as such could be optionally stored also.

[0149] Quantitative characteristics having the meaning of the duplicates (repetitions) number are computed in two steps. At the first step, a set of entries, within which it is necessary to find duplicates, is obtained from the database. For example, in order to calculate the total number of publications duplicating each other and posted on the compromising material aggregators (Ndk), a list of publications posted on the compromising material aggregators is obtained.

[0150] In this example, a list obtained at computing the estimate of the number of publications, made on the compromising material aggregators (Nk), and stored at step (254) could be used. In the other example, in order to calculate the total number of links duplicating each other (Nld), the data obtained in the course of step (240), where the links (URLs) available in the found publications have been extracted and stored could be used. In this case, a query to the database is made by any well-known method, and a list of links found at step (240) is obtained.

[0151] At the second step, the number of duplicates within the resulting list is determined. In order to calculate the total number of publications duplicating each other and posted on the compromising material aggregators (Ndk), this could be performed completely analogously to the step described earlier (142). In order to calculate the number of duplicates in the list of links, a similar algorithm can be used, with the only difference that search in the database is carried out not in the "Publication" field, but in the "Hyperlink" field.

[0152] As shown in Fig. 2B, computing the characteristics having the meaning of the accounts number begins at step (251), where at least one criterion (tag) is selected for filtration, for example, the "Bot" tag. It is selected from the a preliminarily prepared list of tags.

[0153] Then, the method proceeds to step (255), where a query is made to the database comprising the selected tag, and a list of accounts corresponding to this query is obtained from the database.

[0154] Then, at step (256), the obtained list is filtered excluding repetitions from it by any well-known method. Then, at step (257), estimate of this list length, that is, the number of accounts meeting the specified criterion is obtained. Then, at step (258), the obtained estimate as such is stored; and the list of accounts could also be optionally stored.

[0155] For example, in order to determine the number of accounts controlled by bots from which publications have been posted on the compromising material aggregators (Nak) from the list of publications made on the compromising material aggregators. Moreover, the specified list of publications could have been obtained earlier, as described in relation to steps (251) ... (254), or

regenerated by requesting all the publications made by sources tagged as “Compromising Material Aggregator” from the database. Then, a list of accounts, from which they have been made, and having the “Bot” tag, is extracted from this list. The list of accounts is filtered by any well-known method, removing duplicates from it and leaving one entry of each account in the list. The length of the list obtained after such filtration is taken as the desired number of Nak accounts and this number is stored.

[0156] There is a possible embodiment, wherein step (251) is skipped. Thus, in order to determine the total number of accounts from which the found publications (Na) have been posted, a complete list of accounts, from which the publications have been made, is extracted from the database. Then, duplicates are removed from this list by any well-known method, in other words, leaving one entry of each account in it. Number of lines of the obtained list is considered to be the number of accounts from which the found publications Na have been posted, and this number is stored.

[0157] Thus, returning to Fig. 2A, at step (250) the values of the listed characteristics are calculated and stored in the database, then, by any well-known method, for example, by comparing the readings of the system clock at step (220) starting and in real time, the time Tr actually elapsed since the beginning of this iteration is computed, then an estimate of the time elapsed since the beginning of step (210) is computed:

$$T_i = T_i + T_r,$$

[0158] Then, the method proceeds to step (260), where it is checked whether a preset value of the time interval t is reached, by comparing t and Ti. If

$$T_i < t,$$

i.e. the set time interval has not been reached yet, a pause dT is maintained, which is numerically equal to the difference between the preset step size (interval between iterations) ts and time Tr actually elapsed from the beginning of this iteration:

$$dT = t_s - T_r,$$

after that the method returns to step (220), wherein the Internet is crawled and the web pages comprising at least one word or phrase obtained at step (210) characterizing the object of the reputation attack if found.

[0159] If

$$T_i > t,$$

i.e. the preset time interval has been reached, the method proceeds to step (270).

[0160] At step (270), the parameters characterizing a probability of reputation attack based on the calculated values are computed.

[0161] As mentioned before, the values of all quantitative characteristics computed at step (250) are stored in the database in the form of vectors, i.e. a sequences of numbers. For example, as a result of

execution of steps (220) ... (260) within the given time interval t , five values have been calculated for each of the numerical characteristics named in the list (2):

- $N = (N1, N2, N3, N4, N5),$ (3)
- $Nb = (Nb1, Nb2, Nb3, Nb4, Nb5),$
- $Nk = (Nk1, Nk2, Nk3, Nk4, Nk5),$
- $Ng = (Ng1, Ng2, Ng3, Ng4, Ng5),$
- $Ngk = (Ngk1, Ngk2, Ngk3, Ngk4, Ngk5),$
- $Nr = (Nr1, Nr2, Nr3, Nr4, Nr5),$
- $Ngr = (Ngr1, Ngr2, Ngr3, Ngr4, Ngr5),$
- $No = (No1, No2, No3, No4, No5),$
- $Nu = (Nu1, Nu2, Nu3, Nu4, Nu5),$
- $Nh = (Nh1, Nh2, Nh3, Nh4, Nh5),$
- $Nd = (Nd1, Nd2, Nd3, Nd4, Nd5),$
- $Ndk = (Ndk1, Ndk2, Ndk3, Ndk4, Ndk5),$
- $Ndbk = (Ndbk1, Ndbk2, Ndbk3, Ndbk4, Ndbk5),$
- $Nld = (Nld1, Nld2, Nld3, Nld4, Nld5),$
- $Na = (Na1, Na2, Na3, Na4, Na5),$
- $Nab = (Nab1, Nab2, Nab3, Nab4, Nab5),$
- $Nak = (Nak1, Nak2, Nak3, Nak4, Nak5),$
- $Nabk = (Nabk1, Nabk2, Nabk3, Nabk4, Nabk5),$
- $Nar = (Nar1, Nar2, Nar3, Nar4, Nar5),$

[0162] At step (270), an absolute D (in units) and relative Dr (in percent) difference between the adjacent values in each sequence shown in the list is computed (3). For example, for a sequence computed for the total number of publications N :

$$N = (N1, N2, N3, N4, N5),$$

in this example the following will be computed:

$$D1 = N2 - N1;$$

$$Dr1 = 100 * (N2 - N1) / N1,$$

$$D2 = N3 - N2;$$

$$Dr2 = 100 * (N3 - N2) / N2,$$

$$D3 = N4 - N3;$$

$$Dr3 = 100 * (N4 - N3) / N3,$$

$$D4 = N5 - N4;$$

$$Dr4 = 100 * (N5 - N4) / N4.$$

[0163] After computing all the values of the absolute D and relative Dr difference for each sequence of numbers (3) obtained for the quantitative characteristics (2), step (270) ends, and the method proceeds to step (280).

[0164] At step (280) it is determined whether at least one of D and Dr values exceeds a preset threshold value.

[0165] For example, for the numerical characteristic Ndk, which has the meaning of the total number of publications on the compromising material aggregators, being duplicates of each other, there could be set a threshold value of 7 for the absolute difference D, and a threshold value of 5% for the relative difference Dr.

[0166] At the same time, for the numerical characteristic Nr having the meaning of the number of publications made on advertising platforms, there could be set a threshold value of 3 for the absolute difference D, and a threshold value of 6% for the relative difference Dr.

[0167] Moreover, for the numerical characteristics of Nd, which has the meaning of the total number of publications being duplicates of each other, there could be set a threshold value of 95 for the absolute difference D, and a threshold value of 20% for the relative difference Dr.

[0168] In other words, the corresponding threshold values for the relative and absolute difference could be set for each of the named (2) quantitative characteristics.

[0169] These values could be selected empirically at the stage of the system setup.

[0170] If none of D and Dr values is exceeded by the corresponding threshold value, the method returns to step (220), where the Internet is crawled and the web pages comprising at least one word or phrase obtained at step (210) characterizing the object of the reputation attack is found.

[0171] In the other possible embodiment of the described method, in this case the method (200) ends.

[0172] In the other possible embodiment of the described method (not shown in Fig. 2A), the system implementing the method (200) generates a message stating that a reputation attack on a specified target has not been detected and proceeds to waiting for further user commands, for example, entering new words and/or phrases characterizing the object of the reputation attack.

[0173] If at step (280) it is determined that at least one of D and Dr values exceeds a preset threshold value, the method proceeds to step (290).

[0174] At step (290), the estimates of the attack method and the attack nature for different quantitative parameters are computed, based on the calculated values, namely the values of the absolute D and the relative Dr difference. Moreover, at this stage, a notification about reputation attack and also about the method and nature of its implementation is generated and send.

[0175] A non-limiting example of the method for calculating (300) estimates of the attack method and attack nature will be described below with reference to Fig. 3.

[0176] It is worth noting that the algorithm, shown in Fig. 3 as such, is used only for the sake of simplicity of the general illustration of the method; two characteristics Nd (the total number of publications being duplicates of each other) and Nld (the total number of links being duplicates of each other) shown in Figure 3 are also given for ease of illustration and do not limit the method (300).

[0177] All quantitative characteristics given in the list (2) could be used in the embodiment of the method. Moreover, the described method may also include any other, in addition to those shown in Fig. 3, logical dependencies between the characteristics given in the list (2) and be implemented taking into account any predetermined ratios between the numerical values of the quantitative characteristics given in the list (2).

[0178] Similarly, the attack methods shown in Fig. 3, conventionally named as "Seeding" and "Acceleration", do not constitute an exhaustive list of possible methods of reputation attack, and are given by way of example only. The described method allows to identify and reveal, without limitation, any methods of reputation attack known to those skilled in the art.

[0179] The method (300) begins at step (310), where it is determined which quantitative characteristics of those given in the list (2) relate to the values of the absolute D and/or the relative Dr difference that have exceeded a preset threshold.

[0180] For example, if the threshold has been exceeded by the Nld value corresponding to the total number of links being duplicates of each other (320), then at step (340) the "Acceleration" type is assigned to the attack. (This could be a name of the attack, which consists in distributing the same hyperlink to one material influencing the target audience over a large number of web sites).

[0181] Then, the method proceeds to step (360), where the attack level is determined depending on which of the D and Dr values has exceeded the threshold. In this case, if the preset threshold has been exceeded by the absolute difference D value, then, the method proceeds to step (397), where the "Warning" level is assigned to the attack. Otherwise, if the preset threshold has been exceeded by the relative difference Dr value, then, the method proceeds to step (398), where the "Threat" level is assigned to the attack. After that, the method ends.

[0182] If at step (310) it is determined that the threshold has been exceeded by the Nd value corresponding to the total number of publications being duplicates of each other (330), then, at the next step (350) the "Seeding" type is assigned to the attack. (This is a type of attack, which consists in distribution of one and the same text, which content is to influence the target audience, over a large number of web sites).

[0183] Then, the method proceeds to step (370), where the attack level is determined depending on which of the D and Dr values has exceeded the threshold. In this case, if the preset threshold has been exceeded by the absolute difference D value, then, the method proceeds to step (398), where the "Threat" level is assigned to the attack. Otherwise, if the preset threshold has been exceeded by the relative

difference Dr value, then, the method proceeds to step (399), where the highest “Attack” level is assigned to the attack. After that, the method ends.

[0184] It is important that the selection at step (310) is not binary, as shown in Fig. 3 for the sake of simplicity. At this step, any number of characteristics could be selected from those given in the list (2), which corresponding D and/or Dr values have exceeded the threshold. If two, three or more characteristics have been selected, then the sequences of actions corresponding to steps (320) and (330) are executed simultaneously.

[0185] Accordingly, several types could be assigned to such an attack; with reference to Fig. 3, for example, the attack could be simultaneously a "Seeding" and "Acceleration" types.

[0186] Similarly, it is possible a situation when several different levels are assigned to the attack; with reference to Fig. 3, for example, the “Warning” and “Attack” levels could be assigned. In such a situation, the system implementing the described method selects the highest of the assigned levels and uses it when generating an attack notification.

[0187] As follows from Fig. 3, an embodiment of the described method is possible, wherein the reputation attack notification, which is the result of the described system operation, could have one of three severity levels: "Warning", "Threat", "Attack". The said severity levels indicate the level of attack intensity.

[0188] In the other possible embodiment (not shown in Fig. 3) a reputation attack notification could have a numerical expression characterizing the level of attack intensity, for example, “There has been detected an attack on [the attack object name] with the intensity of $I = 71\%$ ”. Moreover, this number I could be obtained, for example, by normalizing the absolute D or relative Dr difference values of any of the characteristics given in the list (2) to the maximum value found over a set time interval t:

$$P = 100\% * (D/D_{\max});$$

[0189] or by any other method based on the numerical values of the quantitative characteristics given in the list (2), for example, on the arithmetic mean values calculated for each of them for a set time interval t, etc.

[0190] Generation and sending of a notification could be performed by at least one of the following methods: by e-mail, by sending an SMS, by sending an MMS, by sending a push notification, by a message in an instant messenger, by creating an API event.

[0191] It is worth noting that use of such a notification tool as API events enables to implement additional integration of the described system with various third-party tools, such as public opinion monitoring platforms, security management platforms, SIEM solutions, etc. Actually, the generation of all the listed notifications, such as emails, SMS, MMS, push notifications, etc. could be performed by any well-known method.

[0192] This completes the described method.

[0193] In the other possible embodiment of the described method (not shown in Fig. 2A), the system implementing the method (200), after generating and sending the notification, switches to waiting for further user commands, for example, entering new words and/or phrases characterizing the object of the reputation attack.

[0194] In the other possible embodiment of the described method (not shown in Fig. 2A), the system implementing the method (200), after generating and sending the notification, returns to step (220) and continues to work according to the algorithm described above.

[0195] Fig. 4 illustrates a schematic diagram of the computer device (400) processing the data required for embodiment of the claimed solution.

[0196] In general, the device (400) comprises such components as: one or more processors (401), at least one random access memory, or memory (402), data storage means (403), input/output interfaces (404), input/output means (405), networking or, what is the same, data communication means (406).

[0197] The device processor (401) executes main computing operations, required for functioning the device (400) or functionality of one or more of its components. The processor (401) runs the required machine-readable commands, contained in the random-access memory (402).

[0198] The memory (402), typically, is in the form of RAM and comprises the necessary program logic ensuring the required functionality.

[0199] The data storage means (403) could be in the form of HDD, SSD, RAID, networked storage, flash-memory, optical drives (CD, DVD, MD, Blue-Ray disks), etc.

[0200] Interfaces (404) are standard means for connection and operation with server side, e.g. USB, RS232, RJ45, LPT, COM, HDMI, PS/2, Lightning, FireWire, etc. Selection of interfaces (404) depends on the specific device (400), which could be a personal computer, mainframe, server cluster, thin client, smartphone, laptop, etc.

[0201] As input/output means (405) keyboard, joystick, display (touch-screen display), projector, touch pad, mouse, trackball, light pen, loudspeakers, microphone, etc. could be used.

[0202] Networking means (406) are selected from a device providing network data receiving and transfer, e.g. Ethernet-card, WLAN/Wi-Fi module, Bluetooth module, BLE module, NFC module, IrDa, RFID module, GSM modem, etc. Making use of the means (406) provides an arrangement of data exchange through wire or wireless data communication channel, e.g. WAN, PAN, LAN, Intranet, Internet, WLAN, WMAN or GSM.

[0203] The components of the device (400) are interconnected by the common data bus (410).

[0204] In conclusion it is worth noting that the data given in the specification are the examples only, which do not limit the scope of this technical solution specified by the claims. It is obvious to a person skilled in the art that other embodiments of this technology compliant with the essence and scope of this technology could exist.

[0205] The exemplary systems and methods illustrated herein could be described in terms of the functional block components. It will be appreciated that such functional blocks could be implemented by any number of hardware and/or software components configured to perform these functions. For example, the system could use various components of an integrated circuit, such as storage elements, processing elements, logic elements, look-up tables, etc., which can perform a variety of functions under control of one or more microprocessors or other control devices. Similarly, the system software elements could be implemented using any programming or scripting language such as C, C ++, C #, Java, JavaScript, VBScript, Macromedia Cold Fusion, COBOL, Microsoft Active Server Pages, assembler language, Perl, PHP, AWK, Python, Visual Basic, stored SQL and PL/SQL procedures, any UNIX shell scripts and Extensible Markup Language (XML), and different algorithms are implemented with any combination of data structures, objects, processes, procedures or other program elements.

[0206] Moreover, the reputation attack detection system could operate on a single computing device, or on several devices interconnected over a network. Additionally, it is worth noting that the system can use any number of conventional techniques for data transmission, signaling, data processing, network management, etc.

[0207] In this context, the devices mean any computing devices software- and hardware-based, for example, such as: personal computers, servers, smartphones, laptops, tablets, etc.

[0208] A processor, microprocessor, computer, PLC (programmable logic controller) or an integrated circuit configured to execute certain commands (instructions, programs) for data processing could be a data processing device. The processor could be multi-core for parallel data processing.

[0209] Memory devices can include, but not limited to, hard disk drives (HDD), flash memory, ROM (read only memory), solid state drives (SSD), etc.

[0210] It is worth noting that the specified device may include any other devices known in the art, for example, such as sensors, input/output devices, display devices (displays), etc. The input/output device could be, but not limited to, for example, a mouse, keyboard, touch pad, stylus, joystick, track-pad, etc.

[0211] The application materials have represented the preferred embodiment of the claimed technical solution, which shall not be used as limiting the other particular embodiments, which are not beyond the claimed scope of protection and are obvious to persons skilled in the art.

CLAIMS

1. A method for detecting reputation attacks, executed by a computing device, the method comprising steps wherein:

at preliminary stage:

- crawling the Internet and finding a publication sources;
- identifying publication sources used for reputation attacks among the publication sources found;
- finding accounts from which entries are posted in the identified publication sources used for reputation attacks;
- identifying accounts controlled by bots among the accounts found; and
- storing the data obtained on the sources used for reputation attacks and on the accounts controlled by bots in a database; and

at operational stage:

- obtaining words and phrases characterizing an object of a reputation attack;
- crawling the Internet and finding publications comprising words and phrases characterizing the object of the reputation attack;
- extracting hyperlinks from the found publications;
- computing quantitative characteristics of the publications and dynamics of their change using data on the sources used for reputation attacks and on the accounts controlled by bots;
- computing parameters characterizing a probability of the reputation attack based on the computed quantitative characteristics, and in response to exceeding a preset threshold value by at least one computed parameter,
 - determining, based on the computed parameters, a type of the attack and a level of the attack; and
 - generating and sending a notification of the reputation attack, and also of the attack type and level.

2. The method of claim 1, wherein at least the following sources relate to the publication sources used for reputation attacks:

- a compromising material aggregators,
- a social networks,
- a data leak aggregators,
- an advertising platforms,
- a groups of related sources,
- an user feedback aggregators,

- a sites for hiring remote workers.
3. The method of claim 2, wherein the groups of related sources include the groups of sources that post identical publications at least a predefined number of times with a publication time difference not exceeding a predefined threshold value.
 4. The method of claim 1, wherein the accounts controlled by bots comprise the accounts that make at least a predefined number of publications within a predefined period of time.
 5. The method of claim 4, wherein the accounts controlled by bots further comprise the accounts that make publications within a set period of time with a frequency exceeding a set threshold value.
 6. The method of claim 1, wherein the quantitative characteristics of publications include at least the following values:
 - total number of publications,
 - number of posts made by bots,
 - number of publications made on compromising material aggregators,
 - number of publications made by groups of related publication sources,
 - number of publications made by groups of related sources which are also compromising material aggregators,
 - number of publications made on advertising platforms,
 - number of publications made on advertising platforms included in the group of related sources,
 - number of publications made on user feedback aggregators,
 - number of publications made on data leak aggregators,
 - number of publications made on sites for hiring remote workers,
 - total number of publications duplicating each other,
 - total number of publications on compromising material aggregators duplicating each other,
 - total number of publications on compromising material aggregators duplicating each other and made by bots,
 - total number of links duplicating each other,
 - number of accounts from which the found publications are posted,
 - number of accounts, controlled by bots, from which the found publications are posted,

- number of accounts from which the publications found on compromising material aggregators are posted,
- number of accounts, controlled by bots, from which the publications are posted on compromising material aggregators,
- number of accounts from which the publications found on advertising platforms are posted.

7. The method of claim 1, wherein the quantitative characteristic change dynamics is computed based on the value of these characteristics computed within a preset time interval with a preset step.

8. The method of claim 1, wherein the parameters characterizing the probability of the reputation attack are computed for each quantitative characteristic as absolute, expressed in units, and relative, expressed as a percentage, difference between adjacent values of this characteristic.

9. The method of claim 1, wherein the reputation attack notification is transmitted by at least one of the following communication methods:

- e-mail,
- SMS,
- MMS,
- push notifications,
- instant messenger messages,
- API events

10. The method of claim 1, wherein the notification of the reputation attack has a numerical expression characterizing a level of the attack intensity.

11. The method of claim 1, wherein the notification of the reputation attack has one of three levels: "Warning", "Threat", "Attack".

12. A method for detecting reputation attacks, executed by a computing device, the method comprising steps:

- obtaining words and phrases characterizing an object of a reputation attack;

- crawling the Internet and finding publications comprising words and phrases characterizing the object of the reputation attack;
- extracting hyperlinks from the found publications;
- computing quantitative characteristics of the publications and dynamics of their change using data on publication sources used for reputation attacks and data on accounts from which entries are posted in the publication sources used for reputation attacks, the accounts being controlled by bots;
- computing parameters characterizing a probability of the reputation attack based on the computed quantitative characteristics, and in response to exceeding a preset threshold value by at least one computed parameter,
 - determining, based on the computed parameters, a type of the attack and a level of the attack; and
 - generating and sending a notification of the reputation attack, and also of the attack type and level.

13. A system for detecting reputation attacks, the system is configured to crawl the Internet and comprising at least:

- a processor,
- a storage device comprising:
 - at least one database,
 - machine-readable instructions that, if executed by the processor, implement the method according to claims 1-12.