Distributed System of Intelligent Content Monitoring Agents

Artem Soboliev [1], Dmytro Lande [1]

[1] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Peremohy Avenue, 37, Kyiv, 03056, Ukraine*

Abstract

The coverage and generalization of large dynamic information flows constantly generated in the space of the Internet requires qualitatively new methods and approaches to the implementation of measures to ensure the completeness, accessibility and reliability of the target information. In the process of content monitoring of the Internet it is important to use tools of distributed global network content monitoring and creation of multiple interfaces between agents who control collection of information from different Internet segments. This paper proposes methods and tools for monitoring the distributed content of social networks, taking into account the constant changes in the availability of certain segments of the Internet. The proposed solutions are used to populate the content monitoring databases of InfoStream and CyberAggregator web resources and social networks. A system of intelligent content monitoring agents based on several servers located in different data centers is offered. The intelligent system of agents interacts with the management and control system, which ensures an appropriate level of fault tolerance, completeness and reliability of the received information.

Keywords

Information resources, social networks, intelligent agent information gathering, distributed content monitoring, intelligent agent system

1. Introduction

Currently, the level of tasks solved by Internet content monitoring systems is constantly growing - from traditional information retrieval tasks to management, design, modeling and forecasting of various processes/events. It is worth noting that the amount of accumulated information is becoming gigantic (Big Data). Consequently, when creating content monitoring systems, it is advisable to take into account the peculiarities of access to certain segments of the Internet. There is a need to find unconventional approaches to the use of information technology and mathematical methods of collecting, processing and analyzing information [1, 3].

Internet resources and social networks have become a convenient and effective means of communication. They provide a huge freedom of action in the information space, which is mostly open and accessible. When used effectively, Internet resources become a powerful source of information for analytical work, open source intelligence (Open-Source Intelligence, OSINT) [3, 4] and at the same time provide an opportunity to obtain strategically important, expert and at the same time publicly available information, in particular security issues that allow to assess the mood of society in a particular information field. Consideration of information from open sources is of great importance for determining the directions of economic, scientific and technical development, as well as for solving problems in the spheres of security and defense [1].

In the global information and technological environment, rapid response and early warning systems for challenges and threats based on monitoring information via the Internet, OSINT systems are being actively improved. Analysis of technological and information problems and functional needs of such systems shows that the use of distributed content monitoring tools (in particular, creation of networks

XXI International Scientific and Practical Conference "Information Technologies and Security" (ITS-2021), December 9, 2021, Kyiv, Ukraine
EMAIL: artem1sobolev@gmail.com (A. Soboliev); dwlande@gmail.com (D. Lande)
ORCID: 0000-0003-4027-042X (A. Soboliev); 0000-0003-3945-1178 (D. Lande)

of information proxies) and a set of interfaces between intelligent information gathering agents (scanning) is important in the process of analyzing information via the Internet. This is due to the following factors:

1. Rapidly growing need for reliable information for management decision-making;
2. The need to take into account the level of accessibility of Internet segments and their features in the formation of content;
3. Different approaches to the possibility of providing information in different segments of global networks, the constant expansion of software interfaces;
4. The need to reduce the load on computing resources when using software agent systems operating in separate segments of global networks;
5. The need for automated management, design, modeling and forecasting, taking into account the distribution of content in the segments of global networks

Popular social networks contain a huge amount of data about people's daily lives and social interactions, so they carefully check every request for information and do not always allow third-party services to use this information. When these information services notice unusual behavior of a client (in particular, a software client, a data collection agent) that makes a request for their data, they immediately block its access and additionally check requests coming from the IP addresses of this client. This requires that the clients, intelligent software agents that collect information for content monitoring systems, provide certain standard behavior in relation to information services inherent in the average user. This achieves the availability of both individual services and entire segments of the Internet.

The number of such agents can be quite large, and information interaction should be provided between them. This will reduce the load on the target information services in global networks, which will also increase their availability.

The purpose of the article is to describe the features of building a system of distributed monitoring of the content of global information networks, taking into account belonging to different segments with the help of intelligent information gathering agents.

2. Presentation of the basic material of the research

Coverage and generalization of large dynamic information flows continuously generated in the Internet space requires qualitatively new methods and approaches to the implementation of measures to ensure monitoring of their content. Specialized content monitoring systems are used for prompt coverage of the information space. Such systems provide:

1. Responsiveness, which cannot be obtained from traditional search engines, where the time of indexing online content can vary from a day to several weeks;
2. Completeness, both in terms of covering sources and providing materials from those sources that are not provided by news aggregators;
3. Application of analytical tools for automated design, modeling and forecasting of processes/events.

A large number of multilingual information resources complicates their use in information and analytical work. When analyzing or collecting such data, there are problems of processing extremely large amounts of data, searching and navigating in dynamic information flows. To solve these problems, such technological concepts as Big Data, Complex Networks, Cloud Computing, Data/Text Mining are used. [1].

Problems in the dynamics and dimensionality of multilingual information resources of global networks require fundamental research in the field of pattern recognition, discrete mathematics, linguistics, digital signal processing, wavelet and fractal analysis. Although in some cases modern development of technology allows us to find the necessary information in networks, there are still unresolved problems of further analytical processing of this information, the allocation of the necessary factual data, determination of trends in the development of certain subject areas, the relationship of objects, events, recognition of significant anomalies, forecasting, etc. Many of the obtained problems are actual issues of semantic processing of ultra-large dynamic text arrays of information.

Today, some attempts to solve these problems in practice determine the success of such projects as search engines Baidu, Yandex, social network monitoring systems Google Keyhole, Brandwatch, CyberAlert, analytical systems Palantir, Centrifuge. One of the suggested approaches to solving of such problems is based on the system of content monitoring of web resources InfoStream [5] and social networks Cyber Aggregator [13]. Modernization and scaling (formation of multilingual full-text databases, modeling of information flows in huge computer networks) in these systems takes into account the software and hardware placement of the "essence" of the segments of the information space, that is, the deployment of a network of information, proxy servers built on the basis of distributed content monitoring, the use of a system of intelligent agents for collecting information.

Let us consider the need for distributed content monitoring tools. It would seem that at the primary level we can use the data available through traditional network search engines hosted on the servers of well-known news integrators. But in this case, there are a number of problems that prevent further serious use of network resources to perform analytical work [7]:

Not all resources are available in the national segment, in particular, there is no access to some foreign sites and some social networks.

Traditional search engines do not always index news posted on deep levels of websites, news is not always indexed by them in time, social networks and special databases posted on the Internet are poorly covered, there is a problem of the Deep Web. In some cases, when the access is not anonymous, websites or social networks involved in information wars can provide distorted information, fakes, to the mainstream users. In some cases, access to information may be denied even if the information has the status of open to all. In addition, requests that satisfy the information needs of analysts, transmitted in an unprotected form, can disclose these needs to an interested party - an information adversary.

To solve these OSINT tasks, it is necessary to use modern integrated systems that are characterized by the following features:

In order to ensure the simultaneous process of obtaining information from social networks without the use of third-party paid services and to control and manage such a system from a single place, it is proposed to introduce teams of agents that allow downloading and exchanging such information with each other and ensure the integrity of the received data and distribute the load among themselves. This will overcome the complexity of distributed content monitoring of information resources on the Internet [8].

Distributed collection of information from websites and social networks using ensembles of intelligent collection agents distributed in a cloud environment that geographically spans different countries. These agents must interact, exchange information, and pass this information to the analytical part of the OSINT system. Information retrieval agents should execute pre-programmed and customized information gathering scenarios, interact with websites, social networks, deep web databases, news aggregators, preferably (if possible) in an anonymous mode. The use of information retrieval agents as the basis of the system of information proxy servers should ensure the completeness of information in case of blocking of individual agents, prevent distortion and duplication of information transmitted to the OSINT system databases. To prevent information leakage, OSINT analysts should use anonymization, masking, VPN, etc. during data extraction and processing.

According to the source [4], OSINT principles are based on the continuous collection of information from public available sources, then analyzed, preparation and timely delivery of the final result to the customer. In order to solve tasks of timely intelligence based on the information received from OSINT is used the result of systematic collection and processing of analyzed publicly available Information. The basis of cybersecurity using the OSINT principle is determined by a number of aspects, including the speed and cost of information obtaining, its volume, quality, reliability, convenience of further use etc. The process of planning and preparation for OSINT management depends on the following factors [3]:

1. Efficiency in information support is achieved by using the method of collecting information from the Internet, user-generated content, hashtags, geo-tags, etc;
2. The relevance, depth, availability and volume of publicly available information makes it possible to find the information necessary for intelligence without the involvement of other specialized intelligence tools;

3. Simplification of data collection processes. OSINT provides the necessary information, eliminating the need to attract unnecessary technical and human resources;
4. Depth of data analysis. As part of the intelligence process, OSINT enables in-depth analysis of publicly available information to make appropriate decisions;
5. Efficiency. Dramatic reduction of time to access information on the Internet. Fast receipt of valuable operational information. The situation that changes rapidly during crises is most fully reflected in the current news;
6. Volumes. The possibility of mass monitoring of certain information sources in order to find targeted content, people and events;
7. Quality. Compared to the reports of special forces, information from open sources is devoid of subjectivity;
8. Reliability;
9. Ease of use. OSINT-data can be easily transferred to any interested authorities, they are open;
10. Cost. Obtaining data on the price in OSINT is minimal.

There is a problem associated with the large amount of information received from social networks and the analysis of this data, assessing the dynamics and susceptibility to constant change. This problem and ways to overcome it today are called Big Data. In this case, it is problematic to implement the functions of collecting, cleaning, storing, searching, accessing, transmitting, analyzing and visualizing such sets as a complete integrity, rather than local fragments. The defining characteristics of big data are the "three V's": Volume (physical volume), velocity (the rate of growth in data transmission and retrieval and the need for high speed of processing and retrieval of results), variety (diversity, the ability to handle different types of structured data). and weak data, structured data at the same time). [13]

2.1. System Functionality

There are problems with processing large volumes of circulating information necessary to search and navigate in dynamic data streams during the collection and analysis of open data from the Internet. Large number of multilingual dynamic information resources, as well as dominance of information noise complicate the search for the necessary information in the operational analysis, and hence the use of open sources in information and analytical work in general. Most of the above problems are topical issues of semantic processing of large dynamic text arrays of information. Nowadays such technological concepts as Big Data, Complex Networks, Cloud Computing, Data/Text Mining are used to solve these problems. In cybersecurity, the Ontology approach is increasingly used to build models of subject areas. [13]

The actual solution to the problem of creating such a corporate system is the simultaneous use of methods and tools for searching, analyzing and aggregating data from information flows. A system of monitoring and analysis of social media, automatic processing of full texts from social networks for a certain period on the topic of "cybersecurity" has been created. Information is scraped from social networks (blogs, various social networks, websites, messengers, etc.) in search mode. Queries (search key phrases in the relevant social network, otherwise an account is required) are read by the software from special configuration tables. Next comes the search and display of records that match the corresponding queries. After that, unique records are written to the server database. Analysis of existing approaches to the aggregation of thematic news has led to the need and possibility of creating a set of tools for monitoring the content of social networks on specific issues, in particular, cybersecurity. The described system includes personalization tools that provide online access to databases, including from mobile devices, for which the possibilities of RSS formats are widely used. The choice of "off-the-shelf" software modules is substantiated, the development tools (scrapers for social networks, tools for generating dynamic RSS-feeds) are described and the results of their integration into the system are presented. [13]

The intelligent agent system interacts with the management and control system, which ensures an appropriate level of fault tolerance, completeness and reliability of the information received. As a basis for control, management, load synchronization and interaction of the intelligent system of data collection agents developed by the authors, the document-oriented database management system

MongoDB is used, which allows storing information about agents, lists of their corresponding network sources, and algorithms of agent behavior. The MongoDB NoSQL database can be easily deployed in most popular operating systems, is undemanding to resources and can withstand heavy loads.

To provide interaction between a group of agents to obtain information, it is necessary to use a database that will allow synchronizing and launching agents according to certain settings of the intelligent system. This database will be used to continuously write and read data. Since in dynamically growing systems, data volumes tend to increase rapidly, you may encounter a problem when the current resources of the machine will not be enough for normal operation.

To solve this problem, scaling is used. Scaling is of two types - horizontal and vertical. Vertical scaling - increasing the power of one machine - adding CPU, RAM, HDD. Horizontal scaling - adding new machines to existing ones and distributing data between them. The first case is the simplest because it does not require additional program settings and any additional database configuration, but its disadvantage is that it is not suitable for distributed intelligent agents, because each server must run a replica of the database and it is easy to connect additional servers if necessary. Therefore, in our case, horizontal scaling will be used, which has the following advantages:

almost infinite scaling (you can include as many machines as you want);

better data security (only if replication is used) - machines can be located in different data centers (if one of them fails, the others will remain).

In this scheme, in addition to sharing, there is segment replication. Let us say a few words about it. All write, delete and update operations go to the primary (primary), and then are written to a special collection oplog, from where they are asynchronously transferred to replicas - repl.1 and repl.2 (secondary). Thus, data duplication occurs. Why is it necessary?

1. Redundancy provides data security - if the master fails, a vote is taken between the replicas and one of them becomes the master.
2. Master and replicas can be located in different data centers - this can be useful if the server is physically damaged (fire in the data center).
3. Replicas can be used to read data more efficiently. For example, there is an application that has a clientele in Europe and the USA. One replica can be placed in the United States and configured so that American clients read data from it. It should be noted that documents to replicas are received with a delay and it is not always possible to immediately find a document re-recorded to a replica. Therefore, this item is an advantage only if the program logic allows reading on replicas.
4. The replica set scheme is often used in serious production applications where data security is important or there is a large number of reads, and the application logic allows reading from replicas.

We will not dwell on this scheme in detail, because it can be a subject of a separate paper.

In order to ensure the simultaneous process of obtaining information from social networks without the use of third-party paid services, as well as to control and manage such a system from a single place, it is proposed to introduce agent teams that allow you to download and exchange data with each other and ensure the integrity of the data received and distribute the load among themselves.

Access control systems in popular social networks are focused on detecting unusual user behavior. In order to avoid such "non-standard behavior", special algorithms for agents' access to such networks were created, which took into account the amount of time spent in these networks, the amount of information collected per request, and the amount of information provided by such a social network agent. It should also be noted that certain attention in such services is paid to the behavior of clients at night (from the region from which the request is made), since during this period the number of requests from clients should be minimized, otherwise such clients are already an object for research.

In the pilot model of the system of intelligent agents for collecting information for distributed interaction ("Nabla" system, ∇), the authors used 3 servers that were geographically located in different data centers, at a great distance from each other (Fig.1):

1. Netherlands;
2. Ukraine;
3. United States of America.

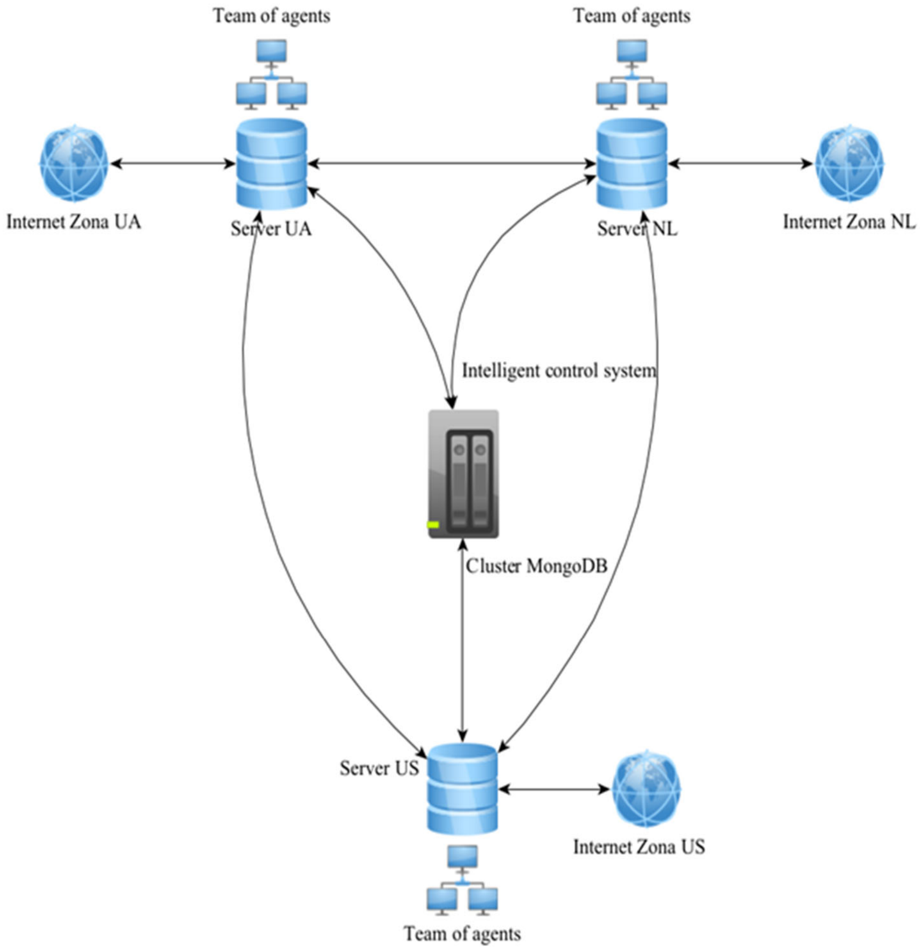

Figure 1: Scheme of distributed information retrieval based on 3 servers

MongoDB database cluster, intelligent management system and agent commands for information retrieval are deployed on these servers. For management and interaction between agents, the HTTPS protocol is used, as it is the most popular protocol on the global Internet, which allows you to quickly optimize commands for network agents and has an appropriate level of security.

Also use RESTful service architecture as the basis for these agent commands, which allows you to configure and monitor their operation efficiently. They also use system messages like Heartbeat as the basis of their interaction, which allows them to efficiently explore the lifecycle of agents and, if any of them fails, quickly identify the problem without losing the data received.

The proposed ∇ teams of agents represent a high availability cluster in which, if one agent fails, its functions are taken over by another available agent. Thus, the process of obtaining information from social networks continues continuously due to the intelligent management and control of these agents. In order to build a fault-tolerant structure, at least two physical servers with storage systems are required, so an auxiliary third server is used to provide fault tolerance on two servers, which allows efficient use of resources and even load balancing between the two servers. Also, the intelligent system of management and control of network agents operates on the following principle: when one agent fails, the other one is automatically included in the work with a message about the agent failure.

A separate agent in the system normally operates in the following scenario (Fig. 2): The agent has a local registry of collected documents, which is constantly synchronized with the general registry stored in the MongoDB DBMS environment. According to the time schedule, the agent selects the task of checking the information resource or part of it, accesses the registers to avoid repeated scanning. Then, if necessary, the information is collected and loaded into the information proxy. Meta-information is written to the local and general registers. The agent then selects a new task again and so on. The process can only end at the command of the system administrator.

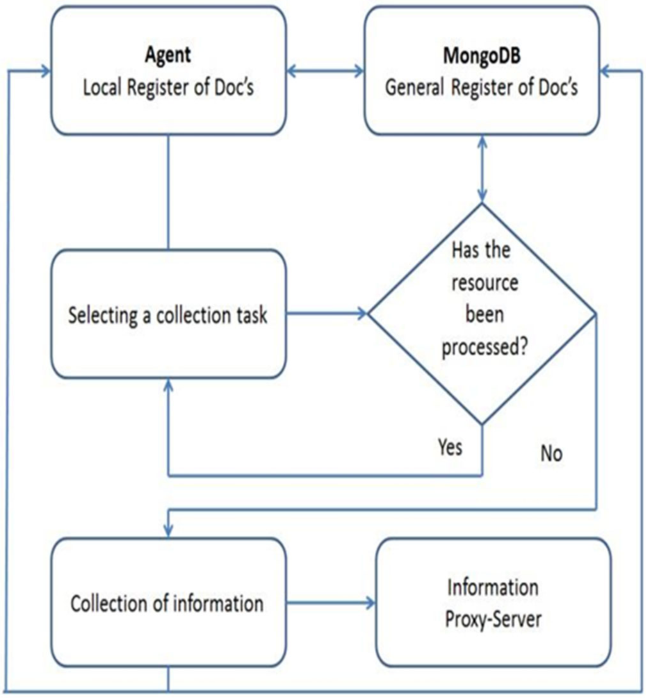
Agent
Local Register of Doc's
MongoDB
General Register of Doc's
Selecting a collection task
Has the resource been processed?
Yes
No
Collection of information
Information Proxy-Server

Figure 2: Scheme of the information collection agent functioning

The general logic of the agent cluster is created at the level of software protocols and allows users to:

1. Manage all network agents with a single intelligent module;
2. Add and update software and hardware resources without system shutdown or major architecture changes;
3. Ensure uninterrupted system operation in case of failure of one or two agents;
4. synchronize data between clusters of agents;
5. Efficiently distribute requests to agent clusters;
6. Use a common database of agents.

One of the servers in the cluster is the central server of the cluster. The central server, in addition to serving client connections, manages the entire cluster and maintains a cluster registry for this purpose.

When a connection is established, the agent communicates with the cluster's central server. The central server, based on the analysis of the agent's workload statistics, directs it to the specific workflow it should perform.

So, the main task of the agent cluster is to eliminate system downtime and report how much the agents collected themselves and how much they took from other agents. Ideally, any incident related to external interference or failure of an internal resource should allow the system to continue working.

In the distributed system under consideration, shared memory, through which different agents can exchange data, is almost never used. Thus, traditional synchronization and communication methods can be considered excluded. It is a set of autonomous agents that are logically combined in communication networks to perform the task of collecting and exchanging data and information about them. With the help of individual agents and MongoDB-based control and management tools, distributed actions are coordinated and information is exchanged.

The considered system of intelligent agents for collecting information, which monitors distributed content, is currently used in the InfoStream and CyberAggregator systems, which distribute monitoring by language (Ukrainian, English, German, Italian, Spanish, French, etc.), by segments of the Internet (web resources, social networks: Youtube, Twitter, Telegram, Facebook, etc.). In this case, information proxies, covering information collected by their respective teams of intelligent agents, process individual segments of the Internet. This approach reduces the load on computing resources by distributing the power of processor modules and global network segments.

Models built using data mining algorithms can be used to make decisions about connecting an additional source. For example, when outliers appear in the time series of the main data stream, a model that identifies such outliers can be used to determine whether it is an error. The model can be built from previously collected data and emergent situations resolved by experts. In addition, if there is feedback from an analyst while receiving data, it can learn, thus adapting to current conditions.

To make a decision on connecting a new source, it is proposed to use methods of intelligent analysis.

3. Intelligent collection of information for analysis

Raw data analysis can be an alternative to collecting all data before performing the analysis. This approach includes the following key steps:

1. Selecting the main data sources;
2. The primary analysis is performed based on information from these data sources;
3. Prioritizing the query for each data source;
4. Can be done based on different criteria;
5. Requesting other data sources, depending on requirements, if information from the basic data sources is not sufficient for the preliminary analysis and/or the probability that the results are correct is low (e.g. outliers and/or jumps are detected in the data);
6. Data processing and evaluation of results performed separately from the main task, if possible close to the data source (preferably at the node where the data are located or in its local network).

Internet solution systems use cloud computing technologies to analyze data and solve problems with computing resources [9]. The cloud provides scalable computing resources and other tools for creating analytical services. However, this approach retains the listed disadvantages. To solve them, Cisco proposed the concept of fog computing [10].

It extends cloud computing closer to the sources. Fog computing completely solves or reduces the impact of a number of common problems of distributed systems:

1. High network latency;
2. Scalability of information sources;
3. Difficulties associated with endpoint mobility;
4. High cost of bandwidth;
5. Large geographical distribution of systems.

Despite the advantages and popularity of the fog computing concept, there are no ready-made solutions for its implementation. This is explained by both the youth of the concept and the high level of abstraction. One of the solutions that corresponds to the concept of fog computing is distributed data analysis based on intelligent agents (actors) [11].

It can be used for both cloud and fog computing. The proposed approach allows splitting data mining algorithms into "pure" functions and executing them on distributed sources.

The data mining algorithm is represented as a sequence of function calls. For their parallel execution, a function is added that allows parallelizing algorithms. For execution in a distributed environment, the data mining algorithm decomposed into functions was compared with the model of intelligent agents (actors). Thus, the distributed data analysis algorithm is represented as a set of agents that exchange messages with the main agent. Intelligent agents (actors) transfer part of the computation to the sources, which improves the analysis performance and reduces the network traffic between the sources and the cloud. However, this approach has some limitations: it does not allow prioritizing data sources and querying data according to their priorities. In addition, the cost of queries from data sources is not taken into account [12].

Consider a simpler system of two agents (Agent 1 and Agent 2), shown in Figure 3, which collect information from an information resource on behalf of two users (User 1 and User 2). The collected information is placed on proxy servers (Proxy 1 and Proxy 2). The information needs of users are represented by query packages (Query 1 and Query 2). If agents make these requests to an information resource (for example, social networks YouTube or Twitter), they may receive several documents R1 and R2, which may coincide - the same documents may satisfy the conditions of different requests.

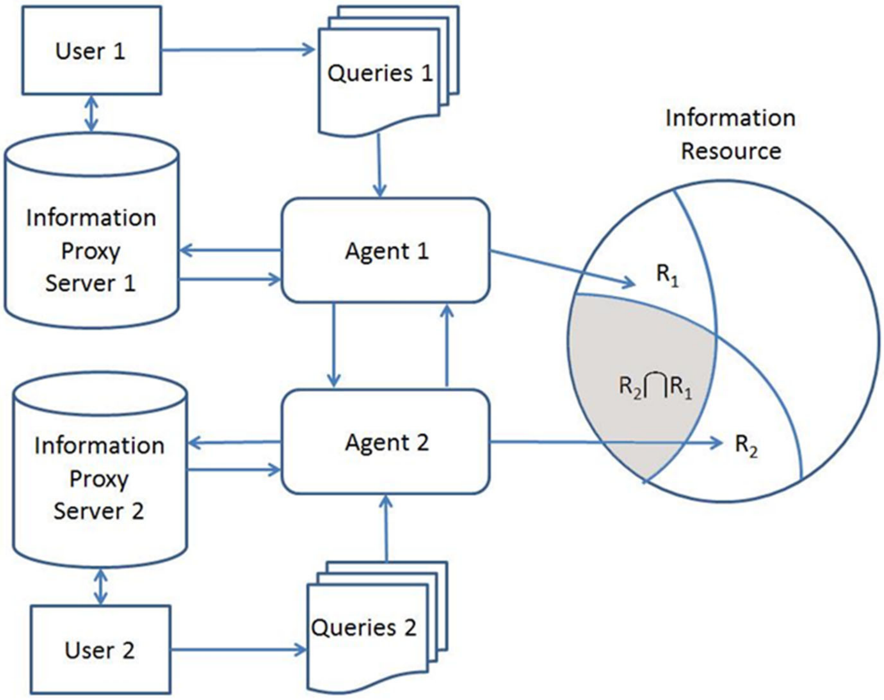

Figure 3: A simpler two-agent information collection system

The rational application of the dual-agent system is to avoid double collection of the same documents due to the interaction of agents. For both request packages, documents already collected by one agent are collected by the second agent not from an external information resource, but from the corresponding information proxy. In this case, the benefit (K) from the application of this scheme of information collection can be calculated as a coefficient:

$$K = \frac{|R1 \vee R2|}{|R1 \cup R2|}, \quad (1)$$

where R1 is resource 1, R2 is resource 2.

4. Conclusions

This paper presents an algorithm for distributed content search agents based on OSINT fundamentals and demonstrates the main possibilities of its use. At the same time, it becomes obvious how diverse and powerful information retrieval processes can be optimized through their interaction. In addition, a mechanism for building a cluster of distributed agents has been presented. It ensures the correct extraction of information. It becomes obvious that the OSINT algorithm is constantly changing and improving, as the owners of open information try to provide a minimum level of access to their resources to other systems.

Based on the testing and evaluation of these network agents, which extract content from the web and social networks and consist of 3 servers, it can be concluded that the proposed interaction is effective. The proposed architecture provided the pilot system with an appropriate level of fault tolerance and reliability of the information received and ensured uniform loading of agent clusters.

In addition, the presented system performs the security tasks of the monitoring service, which allows it to ensure the integrity of the received data, bypassing the limitations in the collection of information, the availability of data for monitoring and the completeness of the received information. If the information for any country is changed, distorted, agents in interaction with each other will reflect these changes and save all copies of the received data.

References

[1] D. V. Lande, Analysis of information flows in global computer networks, Bulletin of the National Academy of Sciences of Ukraine- No 3 (2017), pp. 46-54. doi:10.15407/visn2017.03.045.

[2] H. Liu, A. Gegov, and M. Cocea, Rule Based Systems for Big Data. A Machine Learning Approach. Heidelberg, Germany: Springer, 2016. doi:10.1007/978-3-319-23696-4.

[3] M. Glassman, M. J. Kang, Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT), Computers in Human Behavior, 2012, volume 28, No. 2, pp. 673-682. doi:10.1016/j.chb.2011.11.014.

[4] Army Techniques Publication No. 2-22.9 (FMI 2-22.9), Headquarters Department of the Army, ATP 2-22.9 (Washington, DC, 10 July 2012), URL: https://fas.org/irp/doddir/army/atp2-22-9.pdf.

[5] A. N. Grigoryev, D. V. Lande, S. A. Borodenkov, R. V. Mazurkevich, and V. N. Poter, InfoStream. Monitoring News from the Internet: Technology, System, and Service. Kyiv, Ukraine: Start-98, 2007.

[6] S. Choi, B. Bae, The real-time monitoring system of social big data for disaster management , Computer science and its applications. – Springer, Berlin, Heidelberg, 2015, pp. 809-815. doi: 10.1007/978-3-662-45402-2_115.

[7] A. Hannemann, K. Liiva, R. Klamma, Navigation Support in Evolving Open-Source Communities by a Web-Based Dashboard, IFIP International Conference on Open Source Systems. – Springer, Berlin, Heidelberg, 2014, pp. 11-20. doi:10.1007/978-3-642-55128-4_2.

[8] A. M. Sobolev, D. V. Lande, Distributed Intelligent Content Extraction Agents from Social Networks". Proceedings of the Scientific and Practical Conference "Information and Telecommunication Systems and Technologies and Cybersecurity: New Challenges, New Tasks". - Kyiv: Igor Sikorsky Institute of Cybernetics, 2021, pp. 274-275.

[9] J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, Internet of Things (IoT): A vision, architectural elements, and future directions. Future Generation Computer Systems, 2013, No 29, pp. 1645–1660. doi:10.1016/j.future.2013.01.010.

[10] F. Bonomi, R. Milito, J. Zhu, S. Addepalli, Fog computing and its role in the internet of things. Proc. MCC, Helsinki, Finland, 2012, pp. 13–15. URL:https://conferences.sigcomm.org/sigcomm/2012/paper/mcc/p13.pdf.

[11] I. Kholod, I. Petuhov, N. Kapustin, Creation of data mining cloud service on the actor model. Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Springer, 2015, pp. 585–598. doi:10.1007/978-3-319-23126-6_52.

[12] M. S. Efimova Smart data collection from distributed data sources Software & Systems Received, 2019, vol. 32, no. 4, pp. 565–572. doi:10.15827/0236-235X.128.565-572.

[13] D. Lande; I. Subach; A. Puchkov System of Analysis of Big Data from Social Media Information & Security: An International Journal 47, no. 1 (2020): 44-61. doi:10.11610/isij.4703.

[14] R. Layton, P. Watters, Automating open source intelligence: algorithms for OSINT (Rockland, MA: Syngress Media, 2016), URL: www.bookdepository.com/Automating-Open-Source-Intelligence-Robert-Layton/9780128029169.

[15] B. Akhgar, P. Saskia Bayerl, F. Sampson, Open Source Intelligence Investigation: From Strategy to Implementation (Springer International Publishing AG, 2016), doi:10.1007/978-3-319-47671-1.D.

[16] U. K. Wiil, Counterterrorism and Open Source Intelligence (Wien: SpringerVerlag, 2011), doi:10.1007/978-3-7091-0388-3.

[17] B. J. Jansen, D. L. Booth, A. Spink, Determining the informational, navigational, and transactional intent of Web queries,Information Processing & Management, 2008, volume 44, No 3. – C. 1251-1266. doi: 10.1016/j.ipm.2007.07.015

[18] G. Gutin, T. Mansour, S. Severini, A characterization of horizontal visibility graphs and combinatorics on words, Physica A: Statistical Mechanics and its Applications, 2011, volume 390, No. 12, pp. 2421-2428. doi: 10.1016/j.physa.2011.02.031.