Fact-checking algorithms for the Internet

A S Martynov and I E Voronina

Department of Applied Mathematics, Informatics and Mechanics, Voronezh State University, 1, University Square, Voronezh, 394018, Russia

E-mail: alexander.ser.martynov@gmail.com, irina.voronina@gmail.com

Abstract. Verification of facts refers to the task of verifying textual content. Taking into account the growth of false information on the Internet, fact-checking becomes not only the activity of only journalists. The problem in computer-based fact-checking on the Internet is the lack of universal mathematical models for extracting useful textual information. Proposed the algorithms and mathematical models for extracting useful textual information in order to further verify facts. Presented results of experiments with the practical application of the developed mathematical model for extracting useful text from Internet resources.

1. Introduction

Currently, the Internet is the largest repository of information. Unfortunately, the information posted on Internet resources is far from always reliable, and in the conditions of continuous and constant content updates, it is often difficult to say which sites provide truthful information and which spread outright lies.

Thus, the problem of finding effective large-scale response measures to disinformation on the Internet for journalists, politicians and technology companies has become very important. Dissatisfaction with the so-called "fake news" of political lies and rumors on the Internet has taken on a systemic character. To cope with the dissemination of misleading information, some large web companies have implemented fact-checking services, in which fact-finding information is collected from websites where journalists manually evaluate the quality of the information.

The influence of "fake news" in different parts of the world remains poorly understood [1, 2]. Initial data from the United States and Europe suggests that the proportion of online users who directly visit false news sites is very limited and that people who visit these sites visit major original news sources more often. However, over the past decade, a new phenomenon has appeared on the Internet - social networks. Thanks to social networks such as Facebook, Twitter, Instagram, etc., it is easy to get some false messages from unknown news sources around the world.

A study made by a group of scientists from the Massachusetts Institute of Technology showed that on the social network Twitter, the false news spreads much faster than true information [3]. Based on one hundred and twenty-six thousand publications, which were shared by about three million people on Twitter from 2006 to 2017, it was found that a lie is 70 percent more likely to be reposted - published by another user on a social network with citing a source. In addition, a software robot - a special program that automatically performs human actions, can accelerate the spread of false news. Researchers at the Massachusetts Institute of Technology, using software to identify and filter robots,
found that with or without robots, the results were almost the same. According to scientists, the spread of false information is essentially not connected with computer programs focused on the spread of false information. Instead, according to the authors of the study, people prefer false news and are the main culprits for the distribution of unverified, inaccurate news.

These issues have brought new attention to Automatic Fact Checker (AFC) technology to combat false information on the Internet [4, 5]. Fact-checking - checking the authenticity of the information described in the online resource aimed at identifying discrepancies between the available evidence and the surrounding reality, can partially solve the problem of the detection of false news. New research project Laboratory of Computer Science at the Massachusetts Institute of Technology and Research Institute Qatar aims to develop a system that uses machine learning to determine which sites are based on facts, and which are likely to provide misinformation [6]. The system requires only about 150 articles posted on review sites to determine exactly whether you can trust the source of the information. This system can, will eliminate new fake news, before they are widely distributed.

The scope of fact-checking is not limited to fake news. The ideas proposed by fact-checking can be extended to areas such as marketing, political science, history, and others. In addition, the facts of human life can be confirmed or disproved by validating over biography.

However, the determination of the truth of the facts from misinformation is a complex and often controversial work - these problems become the AFC. Based on reviews of ongoing AFC studies, it was found that:
- verification of facts by humans requires a kind of understanding of the context, which is still far away from fully automatic verification;
- progress has been made in automatically validating for a narrow range of simple factual statements for which reliable data are available. Even then, though, the AFC systems will require human control for the foreseeable future;
- most studies agree that automatic fact-checking is complicated by understanding the natural language context. In particular, difficulties are exacerbated due to increased volumes stored and continuously updated information, and increasing user demands for quality and precision processing. So, according to some estimates, the amount of information on the Internet is growing exponentially every year.

To simplify the search for information, special search engines are used [7], which, in turn, use search robots. Search robot or web spider - a program that is an integral part of the search system and designed to iterate over the Internet in order to enter the information found into the search engine database.

It should be noted, the result issued by the search engine is not always as expected. Often web pages contain a lot of redundant information. In general, a web page consists of many different parts: useful text, somehow marked out, menus, header and footer, links, etc. Links can navigate to other pages or sites, as well as identify resources, such as text files, pictures, scripts, videos, etc.

The proposed work is part of research conducted in the field of verification of the facts, aimed to develop a mathematical model of extraction target text information from Internet resources.

2. Materials and methods

2.1. Goals of the study

The aim of the study is to develop mathematical models and algorithms for extracting target textual information from Internet resources. To achieve this goal it is necessary to solve the following tasks:
- an analysis of existing solutions (crawlers), designed to extract text documents;
- based on the analysis, determine the list of criteria and compare existing solutions and their compliance with the advanced criteria;
- based on the analysis, identify contradictions and identify the problems of existing solutions;
- determine the goal and formulate requirements for the implemented mathematical model;
- analyze and evaluate algorithms, develop a set of utilities and algorithmic tools to extract text information from web pages;
- based on the analysis of requirements, select the means of implementation;
- develop a logical and physical database model;
- perform software implementation;
- analyze the results.

2.2. Mathematics methods and algorithms

For verification of the facts is required information. Earlier studies were conducted in the field of extracting target textual information from Internet resources [8]. At first glance, the task of extracting useful text from web pages from various sites may seem trivial, but its implementation, as a rule, comes down to creating applications targeted at specific sites. To check the reliability of the facts about the person we would like to have a more flexible solution. As you know, a web page consists of many parts: useful text, marked up in some way, menus, headings, etc. However, if delete all the markup it does not work, because there is too much unnecessary text left.

In [9], one of the possible solutions to this problem was considered. The essence of the proposed approach is to split the web page into a large number of parts - it can be lines or paragraphs, and for each part count the number of markups.

Consider an example of the algorithm. For example, we will take an article from the Habr site. The page code was divided into lines and for each line, the number of markups and the amount of text were calculated. The red line on the graph indicates the length of the line, the blue line indicates the number of markups. In order to improve visualization shows data averaged over five rows (Figure 1) The simplest way to determine where is the text and markup where - a select line, wherein the ratio of length to the length of the marking line is below a certain threshold.

![Figure 1. The ratio of the number of markup to the length of the line on the Habr site.](image)

This approach, however, has a drawback: the operation of the algorithm depends on the line break and is not suitable for all web pages. For example, an article from Wikipedia contains a lot of markup between words, and therefore, with some probability, many lines containing real text may be skipped, since there were many markup characters or the text in the line was too short. You must select a new criterion for evaluating the amount of text that minimizes the loss of useful text for such web pages.
Consider the textual representation of a web page (Figure 2):

```html
<html>
  <head>
    <title>Example</title>
  </head>
  <body>
    <h1>Web page example.</h1>
    <p>This page is a sample.</p>
  </body>
</html>
```

Figure 2. An example web page as an XML document.

According to DOM-model document has a hierarchical structure and is represented as a tree (Figure 3), wherein each HTML-tag - a tree node type "element". The tags embedded in it become child nodes. To represent the text, nodes with the type "text" are created. The number of children is unlimited, but the children themselves can have only one parent.

Based on the DOM model, one can construct a connected graph without cycles $G = (V, E)$, where $V$ is a nonempty set of vertices, and $E$ is a set of pairs of vertices called edges. The vertex $v_i$ contains the number of characters related to the markup and text of the node $i$, as well as its children in the DOM tree, and the weights of the edges reflect the distances between the nodes. The distance $d_{i,j}$ between the vertices $i$ and $j$ can be calculated by the formula (1):

$$
\begin{align*}
  d_{i,j} &= \eta \ast \theta \ast \frac{\lambda_i - \lambda_j}{\chi_i - \chi_j}, \ j \in v_i \\
  d_{i,j} &= 0, \ j \notin v_j
\end{align*}
$$

(1)

Here:
- $v_i$ – is the set of children of the node $i$;
- $\chi_i$ and $\chi_j$ – the number of markup characters contained in nodes $i$ and $j$;
- $\lambda_i$ and $\lambda_j$ – the number of characters of the text at the vertices $i$ and $j$;
- $\eta$ – is the ratio of markup to text in node $j$;
- $\theta$ – the fraction of the text of node $j$ in node $i$. 

From (1) it follows that the more text and less markup contained in node j, the lower the cost of going from vertex i. For our example, let’s make a tree according to the formula (Figure 4).

Figure 4. Tree counted distance.

As can be seen from Figure 4, in our example, the subtree of the <body> element has a distance less than that of <head>, which means it is a solution. After calculating each edge weight, the task of extracting meaningful text can be considered as a search for a minimal graph tree [10].

Now that the weight has been calculated for each edge, the task of extracting meaningful text can be considered as a search for a minimal graph tree.

2.3. The use of heuristic algorithms

For the solution to this problem, you can apply the ant algorithm. The first version of the algorithm, proposed by Dr. Marco Dorigo [11, 12] in 1992, was aimed at finding the optimal path in the graph. The idea of this algorithm is based on the behavior of an ant colony, which is capable of performing complex joint work, for example, searching for food that could not be performed by each element of the system separately without external coordination. One of the confirmations of the optimal behavior of ant colonies is the fact that the path they make to food is minimal. Various variations of the ant algorithm are considered in [13].

Imagine the following algorithm:

Step 1. Let \( t = 0 \).
Step 2. Create \( n \) ants (ant).
Step 3. Put on each edge the initial amount of pheromone, which is initialized with a small positive number so that at the initial step the probabilities of transition to the next vertex are not zero.

Step 4. Every ant \( ant_k(1, n) \) moves from \( i \) to \( j \) node; the transition probability is calculated by the formula (2):
\[ P_{i,j}(t) = \frac{\tau_{i,j}(t)^\alpha \left( \frac{1}{d_{i,j}} \right)^\beta}{\sum_{j \in \text{childnodes}} \tau_{i,j}(t)^\alpha \left( \frac{1}{d_{i,j}} \right)^\beta} \]  

(2)

Here:
- \( \tau_{i,j} \) – is the pheromone level;
- \( d_{i,j} \) – is the heuristic distance;
- \( \tau_{i,j} \) – is the pheromone level.

When \( \alpha = 0 \) the choice of the node with the largest amount of text is most likely, that is, the algorithm becomes greedy. At \( \beta = 0 \) the choice occurs only on the basis of the pheromone, which leads to suboptimal solutions. Having passed the edge \((i, j)\), the ant lays on it a certain amount of pheromone.

**Step 5.** Repeat step 4 until each ant reaches a leaf of a tree.

**Step 6.** Update the amount of pheromone, which varies by the formula (3):

\[ \tau_{i,j}(t + 1) = (1 - \rho) * \tau_{i,j}(t) + \sum_{k \in \text{colony}} \frac{Q}{L(t)_{k}} \]  

(3)

Where:
- \( \rho \) – evaporation rate, price of the current solution for \( k \);
- \( Q \) – a parameter having a value of the order of the price of the optimal solution;
- \( L(t)_{k} \) – pheromone laid by \( k \)-th ant, using edge \((i,j)\).

**Step 7.** \( t = t + 1 \), if \( t < T \) (number of iterations), then go to step 4, otherwise go to step 8.

**Step 8.** End of work.

After performing \( t \) operations, the nodes of the source tree will have a certain amount of pheromone. It remains to select the subtree, which has the largest number of pheromones. Vertices that contain pheromone less than the specified threshold value can be cut off since most likely they do not contain meaningful text.

The quality of the obtained solutions largely depends on the tuning parameters in the probabilistic-proportional rule for choosing a path based on the current amount of pheromone and on the parameters of the rules for postponing and evaporating the pheromone. It is possible that the dynamic adaptive adjustment of these parameters may contribute to obtaining the best solutions. For example, apply machine learning algorithms to configure these parameters.

However, the approach presented has a drawback: heuristic distance is used, which means the algorithm will find an approximate solution, that is, in the result tree, with some probability, some nodes containing real text may be skipped if, for some coincidence, there are many markup characters or text in the node was too short.

3. Results and discussion

There are a number of important practical reasons for analyzing algorithms. The same problem can be solved using various algorithms. Analyzing the heuristic algorithm, one needs to get an idea of how accurate the solution of the problem is using this algorithm. The similarity of the expected solution to the one obtained is the main evaluation criterion.

To test the versatility of the algorithm, various types of sites were considered such as information sites, which mainly contain text, electronic encyclopedias, blogs, news feeds, entertainment sites, etc. For the sake of analysis, a group of websites was created consisting of “Lib.ru” and “Samilib.ru”. - Russian-language electronic libraries, the free Wikipedia encyclopedia, the Habr collective blog, and the Pikabu entertainment site.

An example was created for each site, which was compared with the result. Both texts were passed through the shingles algorithm, which showed the percentage of similarity of the text. The result of the text comparison for the parameters \( \alpha=2 \) and \( \beta=4 \) is presented in Table 1.
Table 1. Text Comparison Result.

| Site     | Original page | Prepared Example | Processed Page | Uniqueness Percentage |
|----------|---------------|------------------|----------------|------------------------|
|          | Number of markup characters | Number of characters of text | Number of markup characters | Number of characters of text | Number of markup characters | Number of characters of text |                  |
| Lib.ru   | 2851          | 164310           | 2125           | 155852                 | 2493                      | 155852                      | 0                |
| Samlib.ru| 54481         | 259324           | 46580          | 235958                 | 46780                     | 236123                      | 0                |
| Wikipedia| 152039        | 25021            | 111253         | 14785                  | 135487                    | 16822                       | 0                |
| Habr     | 351345        | 141454           | 6505           | 54483                  | 6505                      | 54483                       | 0                |
| Pikabu   | 301003        | 110385           | 30525          | 7542                   | 277625                    | 98163                       | 100              |

On the basis of data from the table form the histogram (Figure 5), which displays the number of characters of text in these web pages.

![Figure 5. Histogram of the number of characters of text on the pages.](image)

From the histogram, it follows that the algorithm showed an almost exact solution for the sites "Lib.ru", "Samlib.ru" and "Habr." In addition, the percentage of uniqueness of the shingle algorithm for these texts is less than 2%, which means the equivalence of the texts.

Based on the percentage of uniqueness for the Wikipedia site, we can assume that the example is completely contained in the output text document, however, it contains more characters than the example, which means it contains extra elements. For a more accurate Wikipedia site result, you need to configure the parameters $\alpha$ and $\beta$. 

7
It is worth paying attention to the Pikabu website. From the percentage of uniqueness, it can be seen that the example is completely different from the solution. When analyzing this page, it was revealed that the amount of text contained in the main part of the site is much less than in the comments. Based on this, it can be assumed that the algorithm chose not the main part, but comments on it. However, one should ask: if there is more information in the comments, then what is more important - the content itself or the discussion of it by other users?

4. Conclusion
The developed mathematical model for extracting useful text from resources posted on the Internet is universal and can be applied not only in the field of verification of facts but also in other subject areas.

References
[1] Graves L 2018 Understanding the promise and limits of automated fact-checking. http://reutersinstitute.politics.ox.ac.uk/sites/default/files/201802/graves_factsheet_180226%20FINAL.pdf
[2] Nielsen, Rasmus K, and Graves L 2017 “News you don’t believe”: Audience perspectives on fake news Oxford: Reuters Institute https://reutersinstitute.politics.ox.ac.uk/our-research/news-you-dont-believe-audience-perspectives-fake-news.
[3] Soroush V The spread of true and false news online Science 359 1146
[4] Hassan, Naemul, Fatma A, Chengkai L and Mark Tremayne 2017 Toward Automated FactChecking: Detecting Check-Worthy Factual Claims by ClaimBuster Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax https://dl.acm.org/citation.cfm?id=3098131
[5] Karadzhov G, Nakov P, Marquez L, Barron-Cedeno A and Koychev I 2017 Fully Automated Fact Checking Using External Sources http://arxiv.org/abs/1710.00341
[6] Baly R, Karadzhov G, Alexandrov D, Glass J, Nakov P, Predicting Factuality of Reporting and Bias of News Media Sources MIT Computer Science and Artificial Intelligence Laboratory
[7] Segev E 2010 Google and the Digital Divide: The Biases of Online Knowledge Oxford: Chandos Publishing
[8] Martynov A S and Voronina I E 2018 Metasearch system for extracting text information from the Internet Mathematics, information technologies, applications: proceedings of the interuniversity scientific conference of young scientists and students pp78-81. (in Russian)
[9] How to extract useful text from HTML (http://www.interface.ru/home.asp?artId=27160) (in Russian)
[10] Martynov A S and Voronina I E 2018 The algorithm for extracting text information from web pages Applied mathematics and computer science: modern research in the field of natural sciences. Materials of the IV scientific-practical all-Russian conference (school-seminar) of young scientists 2018 pp96-99 (in Russian).
[11] Dorigo M 1992 Optimization, Learning and Natural Algorithms
[12] Thomas S, Dorigo M 2004 Ant Colony Optimization
[13] Barker T, Von Haartman M Ant Colony Optimization www.courses.washington.edu/ inde510/510/AntColonyOptimization3.ppt