Adaptation of information retrieval methods for identifying of destructive informational influence in social networks

Elena Okhapkina\textsuperscript{1,3}, Valentin Okhapkin\textsuperscript{1,4}, Oleg Kazarin\textsuperscript{1,2}

\textsuperscript{1}Institute for Information Sciences and Security Technologies, Russian State University for the Humanities
\textsuperscript{2}Institute of Information Security Issues, Lomonosov Moscow State University
\textsuperscript{3}Laboratory of Network Technologies, Bauman Moscow State Technical University
Moscow, Russian Federation
\textsuperscript{4}Laboratory of Mathematical Modeling of Complex Systems, Vyatka State University
Kirov, Russian Federation
lenaoakhpina@mail.ru, \{vpokhapkin, okaz2005\}@yandex.ru

Abstract—The article describes computational experiment and further research work in the area of identification of destructive information influence in social networks. The problem of distribution of suicidal content via open sources is presented. On the basis of calculations there was made a conclusions about the prospects of using the methods of information retrieval in the task of identification of the destructive informational influence. The procedure of singular value decomposition of the matrix “pattern-message” is considered as a path of the future research.

Keywords—information retrieval; pattern; cosine similarity; singular value decomposition; suicide

I. INTRODUCTION

Active development and widespread integration of information and communications technologies into many spheres of life and human activities have brought both new opportunities and new challenges to person, society and states of the post-industrial civilization. Today, numerous social services, instant messaging services combined with free access to broadband networks including mobile Internet make it possible to bypass state borders while retaining relative anonymity. Therefore, a person with a set of information and communications technologies becomes a free broadcaster of information. Monitoring of the accuracy and quality of such information is rather hard to be implemented or it is carried out formally [1, 2, 3].

Often, such freedom of action in cyberspace is widely used by intruders, for example, for distributing of suicidal content via social networks [1, 2]. Today social network means wide audience mostly composed of young people, a set of tools for moderation of public and private chats, alert tools for target audience, including links to mobile devices and gadgets [4]. The above software and hardware capacities of social networks, on the one hand, ensure the high degree of integration in the everyday life of a young person, and on the other, make a universal interface between the real and virtual personality. Within the Russian-speaking segment of social networks, according to various estimates, there are at least 1000 communities publishing materials about the ways of “voluntary” terminating of life [5]. According to the World Health Organization Europe and Southeast Asia are the regions with the higher than average rate of suicide. In turn, African and American regions, Eastern Mediterranean countries and the Western Pacific keep this rate within the global average value [6]. It is indisputable that suicidal content posted and promoted on various virtual sites of the Internet has destructive capacity for human psyche. This problem is a variant of the destructive informational influence (DII) and one of the threats of the XXI century that requires prompt detection and response from the state control in sphere of telecommunications, information technologies and mass communications [7].

For identification and isolation of the suicidal content source it seems reasonable to use and improve technologies of information retrieval which are widely used by search engines. It is obvious that the identification of DII with the use of keyword causes problem of an extremely large coverage of the social network messages. But the obtained array of messages may include those that do not contain destructive information influences, for example, posts with citations from scientific articles about suicide, media reports or messages condemning suicide. Then, the task of information retrieval for identifying DII of suicidal content in social networks should have a systemic approach which is expressed in the development of marker system and application of methods for quantitative evaluation of the published messages.

II. VECTOR SPACE MODEL FOR ESTIMATION OF THE RELEVANCY OF SOCIAL NETWORK POSTS AND PATTERNS

This study shows the development of methods for identifying destructive informational influence in social networks, published in [8], and reflects experimental work of these methods. For the purpose of the experiment during the analysis of the messages of the Russian-speaking segment of social networks it was formed an array of 2000 messages expressing suicidal context anyway. Table I shows the most typical examples of such messages. Hereinafter the sequence of the analyzed messages will be called a collection. This term
means all published and open individual messages and branches of the dialogues.

TABLE I. EXAMPLES OF SUICIDAL MESSAGES FOUND IN THE RUSSIAN-SPEAKING SEGMENT OF SOCIAL NETWORKS

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Your death is a balm for the eyes.</td>
</tr>
<tr>
<td>2.</td>
<td>I found a place for my future suicide.</td>
</tr>
<tr>
<td>3.</td>
<td>Nobody will notice my death today...</td>
</tr>
<tr>
<td>4.</td>
<td>If you leave, the world without you would be better. There’s another life after death, you’ll like it and everything you need to do is to accept it. If you become one of us. You will find a dream that lasts forever where you, only you are in the lead role.</td>
</tr>
<tr>
<td>5.</td>
<td>I’m leaving you tonight. All of you. Drugs will help with it. I’ve had a deal with it, miraculously survived. But today, I’ll just double the dose, go to bed immediately and feel no pain. Go to bed and never wake up, that’s it, guys. Sometimes love leads to this. Good night to everyone.</td>
</tr>
<tr>
<td>6.</td>
<td>I quickly adjust to a man. I am a naive fool... I wanna die. Who will be my friend until I decide to die? Write me a personal message with the mark “Dying Swan”.</td>
</tr>
<tr>
<td>7.</td>
<td>I’m looking for suicidal friends.</td>
</tr>
<tr>
<td>8.</td>
<td>People, post pictures with gashes.</td>
</tr>
</tbody>
</table>

Note: Due to the international publication the analyzed collection of messages called “Prone to Suicide” was translated into English, including the results of the calculations, however, the numerical experiment performed for messages in Russian.

Hereinafter we will call mentioned collection of messages “Prone to Suicide”. In order to identify the context of the information collection “Prone to Suicide” as a set of messages of suicidal orientation it is necessary to formalize the pattern as a way of reflecting the problem area: suicide.

III. THE FORMALIZATION OF THE PATTERN AND ITS CHARACTERISTICS

Earlier in [8] a pattern was defined as the collection of messages clearly expressing informational influence. We will use this idea to describe the scheme “pattern – message” to assess the relevance of messages posted in social networks to the pattern that characterizes the message as a suicidal one. For the purpose of the experiment such pattern will be called “Suicidal content” (Table II). Later in this work we will develop this approach.

TABLE II. TERMS OF PATTERN “SUICIDAL CONTENT”

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>death</td>
</tr>
<tr>
<td>2.</td>
<td>die</td>
</tr>
<tr>
<td>3.</td>
<td>suicide</td>
</tr>
<tr>
<td>4.</td>
<td>drugs</td>
</tr>
<tr>
<td>5.</td>
<td>jump</td>
</tr>
<tr>
<td>6.</td>
<td>loop</td>
</tr>
<tr>
<td>7.</td>
<td>blade</td>
</tr>
<tr>
<td>8.</td>
<td>vein</td>
</tr>
<tr>
<td>9.</td>
<td>cut</td>
</tr>
<tr>
<td>10.</td>
<td>hopelessness</td>
</tr>
</tbody>
</table>

While analyzing even a small part of the pattern “Suicidal content” you will notice that not all terms clearly expresses suicidal context but in combination with other terms they could characterize a message as suicidal.

IV. TERM FREQUENCY AND WEIGHTING

We now make comparison using previously introduced concepts of document and pattern. We assume that the analyzed document is close to the pattern in that case if one or more terms of the pattern can be found in the document more frequently than in the others. This method of comparison is actively used by Internet search engines when comparing a request without Boolean symbols and documents on the servers. To determine whether the document match the request it is necessary to calculate frequency of each term in the document which are also presented in the request.

To realize described concept we will use weighting scheme which is called the “term frequency”. According to this scheme each term of the pattern has a weight equal to the number of occurrences of the term in the analyzed message. In other words the weight is equal to the number of occurrences of term \( t \) in the message \( d \) and is denoted \( tf_{r,d} \). In the scientific literature a document \( d \) (or any other text document for analysis) which is presented by the set of weights determined by the \( tf \) weights is called the bag of words model. In the model central attention is paid to the number of occurrences of each term into the message while the exact order of the words is ignored. Following all rules of this model the messages “If I kill myself I'll wake up in another place” and “If I wake up in another place, I'll kill myself” are identical. It is clear that messages with the same “bags of words” in large part are close in meaning. It should be noted that in analyzing collection of messages not all words are equally important. For example, the collection of messages about suicide (suicide propaganda) most likely will contain the term “death” in the majority of messages. In this sense, the use of the term “death” as a marker to determine the similarity of a message has little or no discriminating power in determining relevance of message \( d \) and term \( t \) in the pattern. In this case, it is necessary to introduce a mechanism for attenuating the effect of terms that occur too often in the collection. Such a mechanism can be implemented through introduction of a special coefficient which grows with increasing of the term frequency in the collection of messages.

V. INVERSE MESSAGE FREQUENCY

In tasks of information retrieval document frequency \( df_t \) of term \( t \) is widely used. The difference of document frequency
from term frequency is that \( df_t \) takes into account the number of messages in the collection that contains term \( t \). Terms “suicide” and “die” (including the terms for which “die” is a word form) have similar frequencies, while their document frequencies are different for analyzed messages of the “Prone to Suicide” collection. This can be seen in the example shown in Table III and Fig 1.a, 1.b.

TABLE III. COLLECTION FREQUENCY (CF) AND DOCUMENT FREQUENCY (DF) OF THE TERMS FOR “SUICIDAL CONTENT” PATTERN IN “PROME TO SUICIDE” COLLECTION

<table>
<thead>
<tr>
<th>Term</th>
<th>cf</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>suicide</td>
<td>702</td>
<td>659</td>
</tr>
<tr>
<td>die</td>
<td>1078</td>
<td>953</td>
</tr>
<tr>
<td>drug</td>
<td>490</td>
<td>452</td>
</tr>
</tbody>
</table>

Fig. 1.a. The collection frequency of “Suicidal content” pattern terms in the “Prone to Suicide” collection

Fig. 1.b. The document frequency of “Suicidal content” pattern terms in the “Prone to Suicide” collection

At least one third of the collection contains terms shown in the Table III. Obviously, messages that contain the term “suicide” should be more relevant to the pattern that includes term “suicide” while messages that contain the term “die” are less relevant to the “Suicidal content” pattern. Hence the use of document frequency ensures more comparative force when determining the relevance of the terms in the posted message to the terms of the pattern. For the weight correction of the term based on message frequency we will use the concept of the inverse document frequency of the term \( t \), which is defined as follows:

\[
idf_t = \log \frac{N}{df_t},
\]

where \( N \) is the number of messages in the collection.

Analyzing formula (1) it is possible to notice that the value of the inverse document frequency of a rare term is high while for the terms with high frequency it will take a lower value. The conclusion is confirmed by the calculation of the inverse document frequency for terms “suicide” and “die” from the collection “Prone to Suicide”: \( idf_{\text{suicide}} = 0.482 \) and \( idf_{\text{die}} = 0.322 \) respectively. The precise base of the logarithm does not affect the above conclusion about the size of \( idf_t \).

VI. WEIGHTING BASED ON THE COMBINATION OF TERM FREQUENCY AND INVERSE DOCUMENT FREQUENCY

\[
tf - idf_{t,d} = tf_t \times idf_t
\]

Combination of the term frequency and inverse document frequency into a single integral index in order to obtain the weight of each term in each message allows achieving a number of useful properties:

1. The value of the integral index \( tf - idf_{t,d} \) reaches its maximum value, if the term \( t \) of the pattern could be found many times in a small set of messages. Thereby, a large value of \( tf - idf_{t,d} \) acts as a marker of difference from the general mass of messages.
2. The relevance of the terms of the pattern to the analyzed messages is minor in the case when the term \( t \) is either in a small number of messages, or in the vast majority. Index \( tf - idf_{t,d} \) decreases.
3. Integral index aspires to the minimum value, if the term \( t \) is in the majority of messages.

Following the procedure described above, we will perform the weighting of terms in the collection “Prone to Suicide” with the use of terms of “Suicidal content” pattern.

While analyzing the histograms it can be noted that the terms “suicide” and “die” are actively used in about a quarter of message of the collection “Prone to Suicide”. The term “drugs” is used in one third of messages of the collection. It allows indirectly to judge (at the preliminary stage) that in the analyzed segment of the social network contains propaganda and dissemination of drugs and their precursors. Visually assessing the obtained histogram (see Fig. 2) [9], we note that in the collection of messages “Prone to Suicide” the term “die” is used very often, it is expected due to the specific of the analyzed collection.
It is reasonable to assume that the importance of a message with a large number of terms from the pattern is not always higher than the importance of a message with only one term from the same pattern. In studies of information retrieval \cite{10, 11, 12, 13, 14} it is offered variant $tf-idf$ functions. One of the most common weighting scheme is sublinear $tf$ scaling. This scheme involves the use of piecewise function as follows:

$$\text{sublinear} \times \text{tf} = \frac{1 + \log^{+}(t_{f,d})}{\log(T)}$$

Then, we rewrite (2) taking into account (3).

Afterwards we compute $wf-idf$ for the collection “Prone to Suicide” with the use of sublinear term frequency scaling. To assess the relevance of the terms of the document to the pattern we will use vector form. In this case, the components of the vector $\mathbf{v}(d)$ will be weights of the terms of the message $d$ which were discovered in the “Suicidal content” pattern. Figure 3 shows a histogram for terms in “Prone to Suicide” collection. It is necessary to mention that the pre-ranking of messages by the term frequency or by the order of the terms in the messages was not made.

The weight of any term that belongs to the pattern but was not found in the message is zero. It would be reasonable if at the preliminary stage overlap score measure (5) will be introduced. This value determines the relevance of messages $d$ as the sum of the inverse document frequencies of the terms of the pattern which are also included in the analyzed message.

$$\text{Score}(q, d) = \sum_{i \in d} wf-idf_{i,d}$$

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{The result of “tf-idf” weighting for identified terms of messages from the collection (based on “Suicidal content” pattern)}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3.png}
\caption{The result of “wf-idf” weighting for identified terms of messages from the collection (based on “Suicidal content” pattern)}
\end{figure}
Each document will be displayed on a separate axis. Obviously, such method of displaying eliminates the question of the importance of the order of the terms in the messages (see. the example about suicide above). However, we note that there is a problem of the length of the message which is reflected in the fact that the absolute frequencies of the terms in the message and the pattern (in the two compared messages) may not be identical. In order to eliminate this disadvantage it is possible to use a cosine similarity of vector representations of a message \( \vec{v}(d) \) and a pattern \( \vec{v}(p) \) that can be defined as follows:

\[
sim(d, p) = \frac{(\vec{v}(d), \vec{v}(p))}{\|\vec{v}(d)\| \|\vec{v}(p)\|},
\]  

where the numerator represents the dot product of the vectors of a message and a pattern, while the denominator is the product of their Euclidean lengths \( \vec{v}(d) \) and \( \vec{v}(p) \). The need to normalize vectors of messages and patterns with the use of Euclidean metric associated with length-normalization. In this case, (6) can be rewritten in the following form:

\[
sim(d, p) = \left( \frac{\vec{v}(d)}{\|\vec{v}(d)\|}, \frac{\vec{v}(p)}{\|\vec{v}(p)\|} \right) \]  

In the numerical experiment we will demonstrate that the similarity measure (7) is equal to the \( \alpha \) cosine\(^1\) between the vector of a message and the vector of a pattern. To identify a message as more similar to the pattern \( p \) with the use of vector space model terms we will find such a message \( d \), for which the dot product \( (\vec{v}(p), \vec{v}(d)) \) is the largest.

It should be noted that messages which contain no more than one term from the “Suicidal content” pattern have low relevance, despite the high frequency of the term in the message. From the results of numerical calculations it is clear that the use of cosine similarity to assess the relevance of the message to the pattern has the larger value of the dot product if the number of terms and word forms from the pattern which were identified in the message is the larger.

VII. FUTURE WORK AND APPLICATION OF SINGULAR VALUE DECOMPOSITION OF MATRICES IN THE PROBLEM OF INFORMATION RETRIEVAL

The use of cosine similarity of the pattern and the message has important advantages reflected in the presentation of textual information in the quantitative, matrix form and relative ease of computation. However, the assessment of the relevance of the pattern and the matrix “term-message” with thousands of rows and columns implies a significant amount of computation. It might often be the case when terms of the pattern is not found in the messages or could be found rarely, thus matrix “message-term” consists of significant number of zero elements.

In the tasks of information retrieval singular value decomposition (SVD) of the matrices takes a special place. Elements of such matrices are terms frequencies. For the study such matrix is the matrix “pattern-message”. Mathematically it could be represented as follows:

\[
A = U\Sigma V^T,
\]

where \( A_{vm} \) – matrix “pattern-message”, \( U_{vm} \) – matrix whose columns are the orthogonal eigenvectors of \( AA^T \), \( V_{vm} \) – matrix whose columns are the orthogonal eigenvectors of \( A^TA \). Matrix \( \Sigma_{vm} \) is formed from zeros and numbers \( \Sigma_{pi} = \sigma_i = \sqrt{\lambda_i} \), where \( \lambda_i \geq \lambda_{i+1} \) are eigenvalues of matrix \( AA^T \) and are the same as the eigenvalues of \( A^TA \).

The use of singular value decomposition of the matrix “pattern-message” \( A \) for the solution of the problem of identifying destructive informational influence in social networks has a number of important reasons:

- Due to the fact that singular value decomposition uses unitary matrices, it allows obtaining a geometric representation of the matrix transformations “pattern-message” in \( n \)-dimensional space.
- Singular value decomposition of the matrix “pattern-message” \( A \) is stable, in other words, small perturbations of the matrix “pattern-message” correspond to small perturbations of the diagonal matrix \( \Sigma \) used in the decomposition and vice versa.
- On the basis of the matrix \( \Sigma \) you can determine whether the matrix \( A \) is almost singular. In that case, if this is confirmed, the SVD gives you the ability to lower the rank of the matrix \( A \) with the smallest error.

The use of low-rank approximation methods for indexing and retrieval of documents is called latent semantic indexing. It is obvious that the space of open social networks messages is the big data. This requires both effective methods of assessment of similarity of the pattern and analyzed collections of messages and methods that would allow reducing the dimension of the spaces where the estimate is made. In this sense, promising continuation of the research in the area of identifying destructive information influence of the suicidal nature in social media is the use of SVD for the matrix “pattern-message”.

In addition, it seems reasonable to develop a multi-level structure of the pattern. This approach to marking message as a destructive one also covers those messages that inherently express such an effect in a slight degree. So, the inclusion to the pattern of the verb-term “want” (“wanna”) means that the scope of messages for the analysis will include messages expressing, for example, desire to go to a movie or concert. Although the same term “want” can be used in the phrase “I wanna die”. In this sense, the subsequent term acts as a

---

\(^1\) In the literature devoted to information retrieval and evaluation of the relevance of text messages and queries the value \( \text{sim} \left(x_i, x_j\right) \) is called cosine similarity between documents \( x_i \) and \( x_j \).
clarification of the previous term. If we consider the term "want" ("wanna"), as the term of the first level, and the term "die" as a term of the second level, then the concatenation of terms of the first and the second level of the one pattern allows you more accurately to determine the presence or absence of DII in the message. In fact, the creation of multi-level pattern is a movement from the general to the specific, more fine-tuning system for identifying DII in social networks.

CONCLUSION

Active development of information and communications technologies and implementation of virtual services in the XXI century brought both the possibility of rapid exchange of information, and the embodiment of specific threats to the person in the digital space. Countering the distribution of harmful content for the human psyche requires, on the one hand, adaptation of known methods of information retrieval, and, on the other hand, improvement of approaches to the identification and assessment of the destructive informational influence in the virtual space, in particular in social networks.

Computational experiment shows that vector space model to assess the relevance of the query and document proposed in [15] can be successfully applied for the detection suicidal content in area of open messages of social networks for the matrix “pattern-message”. A characteristic feature of this model which was manifested during the computational experiment consists in the change of the weighting scheme. Thus, the use of the sublinear term frequency scaling instead of the inverse document frequency tends to mark the messages that contain the largest possible number of terms of a pattern than those messages that contain one frequent term of the same pattern. The consequence of this is the assumption that different weighting schemes can produce different estimates in the sentiment analysis of the messages. Probably, the presence of diverse vocabulary of a destructive nature expresses more emotional color than the repetition of the same term.

REFERENCES


