RESULTS OF SIMILARITY ANALYSIS OF ONLINE NEWS IN CZECH REPUBLIC

Ondřej Veselý

Abstract:
This paper presents the result of a six-month long press monitoring of major online news publishers in the Czech Republic. News media play an important role in education, but it has been often hypothesized such media is losing its quality under the pressure of new technologies. To help understand this change we focused on the issue of text re-use as a partial quality indicator. To study this issue it is necessary to recognize its occurrence and find its characteristics to reveal the main causes of text re-use. We collected more than 60,000 articles; both regional and national. Then we used a customized tool to process each text to determine lexical similarity for relevant combinations of the news articles. In following analysis we were able to quantify and visualize the rate of occurrence of re-used text as related to their similarity level or other factors. Specifically we found that for 25% of national level news, the majority of its content is reused. A quarter of reused text is provided by Czech News Agency.

Key words: plagiarism, journalism, online news, Czech Republic, text reuse

1 Introduction

The problematic of online plagiarism, text reuse and misconduct concerns many areas of online environment. Although the research is focused mainly in academic integrity in the context of preventing students to create works which fulfill the definition academic misconduct, the most common situation where students are confronted with text reuse is online journalism. The news media has their own code of conduct which forbids plagiarism, but the trade with news articles is common. Usually media takes over the text from a news agency which leads to situation where the same text is published in many online news servers only with minor modifications. But the situation is complicated: “Lack of attribution and plagiarism can create a special problem for journalists. As numerous examples indicate, there is confusion about the sometimes fine line between lack of attribution and plagiarism.” (White, 1989)

In Czech Republic, there is only one Czech non-specialised news agency called Česká tisková kancelář (ČTK, 2015).

In the context of preventing academic plagiarism it is important to clearly communicate the difference between newspaper articles and the student’s academic or school work. Thank to project Impact of Policies for Plagiarism in Higher Education across Europe (IPPHEAHE, 2015) we have a lots of data about text reuse in academia but there is no precise data about the situation in news journalism. We start with the fact, that “it is known that more than 10–20% of articles collected by portal sites are nearly identical or quite similar” (Chang-Keon R., 2009)

Let Chang-Keon Ryu’s statement be the starting hypothesis we try to clarify.
2 Methods

The results have been created in three steps
1. dataset creation
2. plain text data extraction
3. similarity analysis
4. data visualisation

2.1 Dataset creation

The dataset is a collection of news articles. The articles were collected by method called web scraping. Since the beginning of year 2014, every 10 minutes each of the selected news servers has been checked for new textual content via RSS protocol. If new content is detected, it will be time-stamped and stored in the database. The dataset of national articles consists of 64,806 articles, the regional dataset is based on 7,636 articles.

The main criteria for news server selection was the existence of RSS interface, the second one was it’s traffic listed in Netmonitor service (SIPR, 2015).

2.2 Plain text data extraction

Every article has been converted from HTML format to a plain text. Usually, not all text on the page is part of the article content (for example textual advertisements), these parts have been detected and stripped.

A random inspection of the functionality has been manually performed for some of the stripped text for each monitored server. In is possible that in rare cases not all of the text from the scraped article are used and saved for analysis.

2.3 Similarity analysis

Every day each new article has been compared with other articles within a one-week time window. The length of the windows is limited because of computational performance. But as it can be seen on fig. 1, the chosen window size is big enough to detect almost all similarity occurrences, most of the similar articles are released within the same hour and more than 99% of cases happen within 48 hours.

The basic algorithm for lexical similarity has been used. Every similarity analysis of a pair of the articles A and B produces two values

1. how A is similar to B \( \iff s_{AB} = \text{sim}(A, B) \)
2. how B is similar to A \( \iff s_{BA} = \text{sim}(B, A) \)

For example for texts A = “aaabbb” and B = “aaa”, the results are

\[ s_{AB} = \text{sim}(A, B) = 50\% \]
\[ s_{BA} = \text{sim}(B, A) = 100\% \]

For the purposes of this article the direction of text reuse is not as relevant as the rate of occurrence of the reuse: the higher of this pair of values is used.

A random inspection of the functionality has been manually performed for some of the similarity text pairs for each monitored server via development interface on fig. 2.
PAPERS—SECTION II  Results of similarity analysis of online news in Czech Republic

Figure 1. Relation of number of similar articles based on their published time difference (please note the logarithmic scale).

2.4 Data aggregation and visualisation

8 similarity classes have been recognised (sim() over 20% to 90%). For each class the number of article pairs with corresponding similarity level have been counted (section 3.2).

Also the number of similar articles between different kinds of monitored servers has been visualised to find out the text reuse characteristics (section 3.1)
Results

3.1 Content similarity between individual news servers

Two circle multi-relational graphs show the rate of text-reuse between monitored servers. Two sets of servers were used

1. regional – news servers focused on South Moravian Region of Czech republic; also one blog and two public reporting services (police and firemen) has been selected for mutual analysis

2. national – news servers focused on national news coverage

The relation values are the sum of the articles with similarity level higher than 50% published between September 4th 2014 and March 3rd 2015. In both sets the articles which declare that its original source is ČTK have been excluded from dataset.

Except common regional news servers, some other content sources have been added to visualise its interconnections. Their specifics are described in Table 1.

Heavy text reuse was discovered between a specialised news server just for fire events Pozary.cz and municipal police website for press releases Policie.cz. It can be deduced that these servers have a common owner or both servers have some kind of agreement.


Table 1
List and description of specific content sources from graph on fig. 3

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Žít Brno</td>
<td>Opinion journalism magazine, political satire</td>
</tr>
<tr>
<td>Firebrno.cz</td>
<td>Firemen website for press releases</td>
</tr>
<tr>
<td>Pozary.cz</td>
<td>Specialised news server just for fire events</td>
</tr>
<tr>
<td>Policie.cz</td>
<td>Municipal police website for press releases</td>
</tr>
<tr>
<td>Qrulant blg</td>
<td>An example of common blog focused on regional news</td>
</tr>
</tbody>
</table>

Another interesting pattern was found at the server Brněnská drbna which is connected to all other content sources. Closer examination showed the most probable cause; the server publishes most press releases without any change.

As can be seen from the fig. 4, the Super.cz server appears to have no text reuse activity detected. The reason for this regularity is that Super.cz is not a primary news server, but more like entraining tabloid completely producing its own content.

From the other relations the focus of each server can be deduced on production of its own content. For example Novinky.cz is the most visited news portal in Czech Republic.
Figure 5. Overall similarity on regional news servers dataset shows the number similar articles for each similarity class. The lighter columns show the overall similarity values only for articles which had not been produced by ČTK (SPIR, 2015) known for its low cost content production in large quantities. The graph confirms this general impression.

3.2 Overall content similarity analysis

9 similarity classes were recognised (10% to 90%). For each class the number of article pairs with corresponding similarity level were counted. Also the set of articles originally made by ČTK were separated from the other articles.

Unlike the graphs from section 3.1 these were generated from articles published since 1st January 2014 and news servers were included exclusively.

The fig. 5 shows that the influence of Czech news agency to text reuse is strong, especially for text with highly similarity degree. In comparison with overall similarity on national level, the ratio of it is much lower.

It is estimated that the amount of very similar content in regional media is about five times lower in comparison with national level.

The fig. 6 shows that the influence of Czech news agency to text reuse is weak, especially for text with highly similarity degree in comparison with regional articles. (graph 4)
On the fig. 5 and 6, we can see that the amount of similar content in national media is about five times higher in comparison with regional level.

4 Discussion

The dataset offers a lot of data analysis opportunities, for example detection of most reused text or topic funnel visualisation. Also the method of similarity evaluation could be more semantic-based than lexical to detect the text modification while keeping the same meaning.

Another improvement would be a separation of international news agencies the same way as it has been done with ČTK. This step would quantify the other sources of texts the media uses.

The results can be used in Academic Integrity lessons to illustrate on hard data the different attitude of journalists and academics to text reuse.

The monitoring system is going to be improved to provide automatic validation of news authorship and therefore monitoring of journalism code of conduct compliance. In this context the cooperation with Open Society Fund (OSF Prague 2015) has been established to include press news into the dataset for further analysis. The OSF Prague also nominated results in this paper for Journalism Award, which objective is “to help
improve journalistic professionalism and quality. For this reason, the award emphasizes investigative journalism, which acts as a kind of watchdog of democracy.” (OSF Prague 2015). Eventually some other servers known for spreading propaganda were added into long term monitoring.

5 Conclusion

From the graph 5 it could be concluded that approximately 10% of all articles are 80% or more similar to another one published by different server which makes Chang-Keon Ryu’s statement valid in Czech environment considering the national dataset.

The ČTK is responsible for 20–40% of text-reuse cases on national level, but other causes have not been revealed. On the regional level the ratio of similar text is five times lower and the most similar articles are created by ČTK.

Literature


Copyright statement

Copyright © 2015. Author(s) listed on the first page of article: The author(s) grants to the organizers of the conference “Plagiarism across Europe and beyond 2015” and educational non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to Mendel University in Brno, Czech Republic, to publish this document in full on the World Wide Web (prime sites and mirrors) on flash memory drive and in printed form within the conference proceedings. Any other usage is prohibited without the express permission of the author(s).

Author

Ondřej Veselý (xvesely5@node.mendelu.cz), Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic