International Workshop on Visual Analytics (2012)
K. Matkovic and G. Santucci (Editors)

The News Auditor: Visual Exploration of Clusters of Stories

Michael Behrisch, Miloš Krstajić, Tobias Schreck, Daniel A. Keim

Data Analysis and Visualization Group, University of Konstanz, Germany

Abstract

In recent years, the quantity of content generated by news agencies and blogs is constantly growing, making it difficult for readers to process and understand this overwhelming amount of data. Online news aggregators present clusters of similar stories in a simple, list-based manner, where the most important article is shown first, while all the other similar articles appear below as hyperlinked headlines. This layout makes the user unaware of the content differences between articles, thus making it very difficult to get a comprehensive picture. Understanding what was changed, how, when and by whom, would lead to new insights about the content distribution over the internet and help in dealing with the news overload problem. We present a visual analytics tool that allows the user to compare the articles that belong to the same story and understand the differences at three levels of detail. Story matrix provides an overview of a document cluster, where the user can identify articles of interest based on their overall similarity and reorder them by different criteria. Structural view shows document thumbnails with highlighted paragraphs of the text that were copied, modified or repositioned by different sources. Finally, Document level view presents two articles side by side to provide full-text comparison. To evaluate our tool, we present two user scenarios applied on a real world data set.

Categories and Subject Descriptors (according to ACM CCS): D.2.2 [Software Engineering]: Design Tools and Techniques—User interfaces H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval Design Tools and Techniques—Information filtering

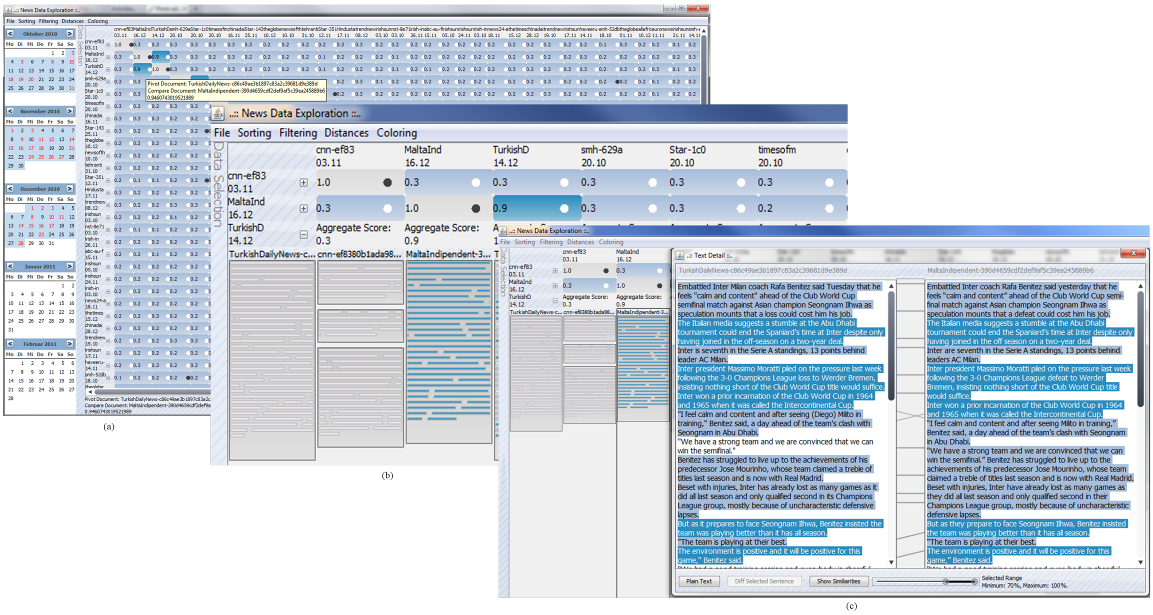
1. Introduction

Websites of newspapers, magazines, radio and television broadcasters publish stories, which are often provided by major news agencies, such as Associated Press, Reuters, AFP, etc. These news stories and feature articles can be prepared by agencies in a way that requires little modification, but very often the clients edit the text before delivering it to the reader. Alternative information flow is created by independent and local media, who publish news stories that are later picked by other media providers and redistributed through their channels.

News aggregators, such as Google News, Yahoo News, or Europe Media Monitor provide the end users with the latest and the most important story clusters, where news articles are grouped by their similarity. The aggregators present the user a simple list, where the first (and the most important) article is usually presented with a title, short summary and a photo, while the other articles are represented as hyperlinked headlines. Navigating through a story cluster becomes a daunting task, since it gets very hard to understand the differences across different sources and find new information without reading every article. A similar problem exists when important breaking events happen, when immediate response by news providers is required. These developing stories are continuously updated as soon as the new information becomes available and the reader needs a fast and effective solution to differentiate the new from the old.

In this paper, we present a visual analytics tool that helps the reader in exploration of a news story cluster. Our work presents a proof of concept that identifies *what* is different in similar news items, combining existing automated methods for measuring text similarity and interactive document visualization. The architecture of the tool allows easy integration of more sophisticated natural language processing methods, which would help the reader in understanding *how* the content is different.

Section 2 describes related work in the visual analysis of news data. Section 3 outlines the problem and the challenges of automatically analyzing and representing online text data. In Section 4, we present our proposed system, while Section



Pivot Document: TurkishDailyNews-c86c49ae3b1897c83a2c39681d9e389d
Compare Document: MaltaIndipendent-390d4659cdf2def9af5c39ea245889b6
0.9460743019521989

TurkishDailyNews-c86c49ae3b1897c83a2c39681d9e389d

Embattled Inter Milan coach Rafa Benitez said Tuesday that he feels "calm and content" ahead of the Club World Cup semifinal match against Asian champion Seongnam Ilhwa as speculation mounts that a loss could cost him his job.
The Italian media suggests a stumble at the Abu Dhabi tournament could end the Spaniard's time at Inter despite only having joined in the off-season on a two-year deal.
Inter is seventh in the Serie A standings, 13 points behind leader AC Milan.
Inter president Massimo Moratti piled on the pressure last week following the 3-0 Champions League loss to Werder Bremen, insisting nothing short of the Club World Cup title would suffice.
Inter won a prior incarnation of the Club World Cup in 1964 and 1965 when it was called the Intercontinental Cup.
"I feel calm and content and after seeing (Diego) Milito in training," Benitez said, a day ahead of the team's clash with Seongnam in Abu Dhabi.
"We have a strong team and we are convinced that we can win the semifinal."
Benitez has struggled to live up to the achievements of his predecessor Jose Mourinho, whose team claimed a treble of titles last season and is now with Real Madrid.
Beset with injuries, Inter has already lost as many games as it did all last season and only qualified second in its Champions League group, mostly because of uncharacteristic defensive lapses.
But as it prepares to face Seongnam Ilhwa, Benitez insisted the team was playing better than it has all season.
"The team is playing at their best.
The environment is positive and it will be positive for this game," Benitez said.

MaltaIndipendent-390d4659cdf2def9af5c39ea245889b6

Embattled Inter coach Rafa Benitez said yesterday that he feels "calm and content" ahead of the Club World Cup semifinal match against Asian champion Seongnam Ilhwa as speculation mounts that a defeat could cost him his job.
The Italian media suggests a stumble at the Abu Dhabi tournament could end the Spaniard's time at Inter despite only having joined in the off season on a two-year deal.
Inter are seventh in the Serie A standings, 13 points behind leaders AC Milan.
Inter president Massimo Moratti piled on the pressure last week following the 3-0 Champions League defeat to Werder Bremen, insisting nothing short of the Club World Cup title would suffice.
Inter won a prior incarnation of the Club World Cup in 1964 and 1965 when it was called the Intercontinental Cup.
"I feel calm and content and after seeing Milito in training," Benitez said, a day ahead of the team's clash with Seongnam in Abu Dhabi.
"We have a strong team and we are convinced that we can win the semifinal." Benitez has struggled to live up to the achievements of his predecessor Jose Mourinho, whose team claimed a treble of titles last season and is now with Real Madrid.
Beset with injuries, Inter have already lost as many games as they did all last season and only qualified second in their Champions League group, mostly because of uncharacteristic defensive lapses.
But as they prepare to face Seongnam Ilhwa, Benitez insisted the team was playing better than it has all season.
"The team is playing at their best.
The environment is positive and it will be positive for this game," Benitez said.

Selected Range
Minimum: 70%, Maximum: 100%.

Figure 1: *News Auditor: The user explores a news story cluster by identifying interesting patterns in the similarity matrix Overview (a); The Structural View (b) provides a visual comparison of differences between selected articles on the paragraph level; The Document View (c) shows direct changes between two articles on the word and sentence level.*

5 demonstrates an application on a real world news data set in two different scenarios. Finally, we describe future work in Section 6.

2. Related Work

In recent years, visualization of text data has been gaining increased interest by researchers, who are developing techniques for efficient display of document collections, as well as single documents. For example, the visual analytics tool *VISRA* [OSSK10] combines readability feature selection with document visualization techniques based on *Literature Fingerprinting* [KO07], *TileBars* [Hea95] and *Seesoft* [ESS92] to evaluate the readability of the input text. In the domain of news analysis, several tools exist that deal with summarization and visualization of news content. Newsmap [Wes12] is a well-known treemap visualization of data gathered by Google News. Other popular news aggregators include Yahoo! News and Europe Media Monitor [EMM12]. The TextMap website, based on Lydia [LKS05], is an entity search engine, which provides information about people and places extracted from the news sources. These systems have limited visualization capabilities that would allow the user to understand the content differences among different sources that provide news reports on the same real world event. In the area of knowledge discovery and data mining, ongoing research efforts exist that deal with meme-tracking [LBK09] and refining causality [SFDBC11]. In this field, the main goal is to find out how the information propagates through networks and how network processes cause a specific behavior in the network, by analyzing appearance of short phrases in document nodes. Researchers working on web indexing and crawling have developed methods for identifying near duplicates [MJS07], i.e. redundant web documents that differ only in a small portion.

3. Problem Description

The goal of our work is to help the user to understand the content of a large document corpus, while understanding the main themes and the various differences among individual news articles. In a real-world scenario, news clusters can contain hundreds of related news articles, but only rarely more than 100 documents per cluster can be retrieved without taking topic shifts or -drifts into account. In this document corpus user should be able to:

- identify interesting articles in the story cluster
- understand *what* are the differences between news articles
- understand *who* changed the content.

Given a cluster of news stories, we assume that the documents are related and can appear in one of two scenarios: a) the documents are news reports from multiple sources on the same event; b) the documents are updates from a single news source on an ongoing event. In order to help the user to get a better insight from a cluster of news articles, we need to combine automated methods for efficient computation of document similarity and visualization techniques that would show the changes at different levels of detail. The algorithmic and visualization challenges and our design decisions are described in the next Section.

4. Design Rationale

We design our system following the overview and detail concept [Shn96], to allow the user exploration of a document collection on different abstraction levels. On the overview level, inter-document distance scores guide the user to interesting patterns within the text collection. A more detailed comparison on the structural level shows the differences between the documents on the paragraph level. Lastly, a document level view shows two articles side-by-side, to provide direct comparison of the texts. Due to this structured approach, it becomes possible to lead the user to non-obvious patterns in a topic-coherent news cluster.

4.1. Overview

The overview visualization, depicted in Figure 1 (a), represents a heatmap color-coded *similarity matrix* and functions as an inter-document comparison view. In the matrix, each cell represents the similarity between the *pivot document* (row ID), and the *comparison document* (header ID). A logarithmic color-to-distance mapping is implemented to emphasize important distance intervals. To guide the user, each cell contains a small black or white glyph that depicts whether the articles stem from the same news source (black dot) or not (white dot). A binning-based or continuous heatmap color coding is used in all aggregation views. The binning-based color codings differ in the number of bins and the base colors. In Figure 1, a light-to-dark-blue color coding with three classes, extracted from [Bre12], is shown. Furthermore, the users can decide to filter out news updates from the similarity matrix.

The matrix view is enhanced by three information filtering and interaction subcomponents, which help in finding patterns of interest. On the left side, a calendar component is used to filter time intervals. The user can control sorting, filtering and coloring settings and choose from three available distance measures, such as Cosine, Google NGD [CV07], or a semantics-driven bag-of-synsets distance. The matrix can be reordered by highest/lowest similarity or according to different usage-driven scenarios that can rely on the article metadata, such as finding copied or reused texts from different news providers, by time of publication, etc. Additionally, the articles can be grouped visually by the news source, showing the update processes happening during the news evolution.

4.2. Structural View

After getting an overview of the news cluster, the user can choose to expand one or more rows to explore the structural features of the documents and their differences. As an example, the rows with very high overall similarity scores in a few or all documents can be regarded as suspects for plagiarism. Structural View is shown in Figure 1 (b). Here, *document thumbnails*, following the visual presentation of [ESS92], visually encode the sentence- and paragraph structure, as well as their textual similarity in comparison to the row's pivot document. The thumbnail width is fixed to allow a comparison of the news articles' text length. Likewise, the sentence bar's length corresponds to the amount of characters in this sentence. The paragraph boxes are determined by the amount of space required by all paragraph sentences, thus leading to a bottom-up layout approach of the *document thumbnails*. The coloring stays consistent with the overview, thus justifying the overview's aggregate scores.

4.3. Text View

For in-depth investigation, the users can switch from the structural view to the textual representation. This component is shown in Figure 1 (c). The text detail view shows the pivot and comparison text in the left and right text panel, respectively. Besides comparing the text by reading both articles, the user is supported by the color coding. The range slider on the bottom of the screen helps the user to highlight sentences within user-selected similarity intervals. Thus, it is possible to filter out all sentences above 80%, see the minimum or the maximum similarity boundaries. Highly similar sentences (above 70% similarity) are visually connected by reference-lines that appear in the space between the documents. By clicking on a sentence, the most similar sentence in the other document is highlighted, showing the word-based similarities with the help of the *Diff* algorithm [HM76], which visually marks insertions/deletions.

5. Use Cases

Various kinds of questions regarding the publication process of news can be answered with our tool. Two use cases are presented in this section to demonstrate the analysis process with the real world data. Our tool uses the data provided by Europe Media Monitor [KMS*10], a publicly available online news aggregator.

5.1. Reuse of Text by different News Agencies

One primary question, which can be answered with *News Auditor*, refers to the reuse and copying of text. In Figure 1 (a), one can see an example for the copying of news from an earlier news source. These are Champions League soccer news articles, which appeared in the period from October, 1^{st} of 2010 until December, 31^{st} of 2010.

The overview is configured with the distance-aggregate sorting option, filtered updates, and the cosine similarity as a text similarity measure. With *News Auditor* most uninteresting documents can be discarded immediately in the Overview matrix due to its low inter-document similarity score (rendered in light blue). Every document with a high similarity score, depicted by a dark blue color, and a later publishing date appears interesting. These characteristics occur, e.g., at the second column and third row. The copying hypothesis is even more obvious if it is not the same source that published the article. In Figure 1, the initial article was published on October, 14th of 2010 from the *Turkish Daily News* agency and modified on October, 16th of 2010 from the *Malta Independent Press*.

The structural comparison in Figure 1 (b) shows that most sentences are in high similarity classes. The structure appears to be stable, yet the length has changed marginally. In fact, the textual investigation, shown in Figure 1 (c), reveals that 21 of 31 sentences are in the similarity interval of 90% to 100% with insignificant changes, such as inserting/deletion of hyphens, quotation marks, or punctuations. Eight sentences have minor modifications, such as plural/singular changes, with a similarity score between 80% and 89%, two sentences are in the 70% to 79% range with word (-suffix, -prefix) exchanges or additions, and only one sentence is in remaining range of 0% to 69%, which has been deleted in the latter news text.

5.2. Updating of News from the News Producer

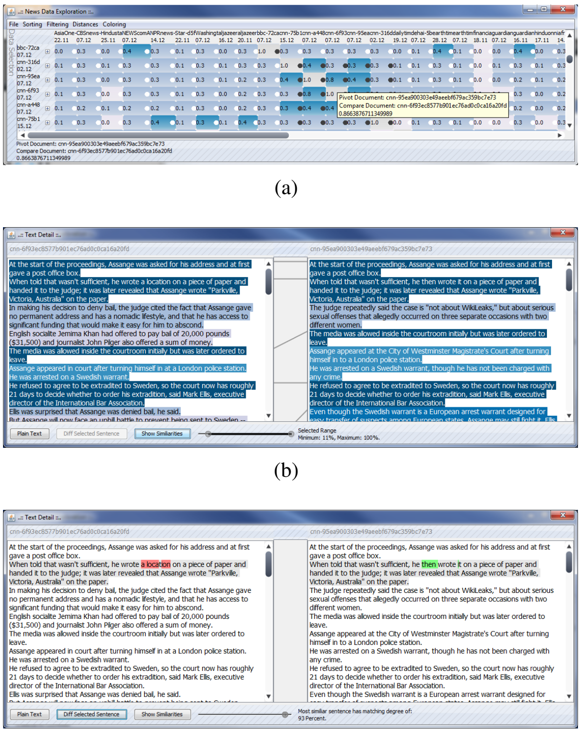
Figure 2 depicts a different use case. Here, the task is to find updates, which stem from the same news source, and compare them with regards to their content. Thus, the similarity matrix is sorted according to the same-sources-first option, without filtering updates, and the Cosine similarity measure. For this specific task, a user needs to find cells that are labeled with a black dot (depicting the same source) and a high inter-document comparison score (depicted by a dark blue color).

As Figure 2(a) shows, a news article by *CNN* can be found in a news cluster that deals with the Wikileaks founder *Julian Assange*. It has been published and modified on December, 7th 2010. Figure 2 depicts in (b) that various modifications have been made to the news article, both in the structure and the text. Despite the case that the majority of sentence are the same, it can be seen that sentences like, e.g., *"English socialite Jemima Khan had offered to pay bail of 20,000 pounds ($31,500) and journalist John Pilger also offered a sum of money."* have been deleted. Figure 2 (c) shows one of the minor textual modifications. Here, *"[. . .] he wrote a location [. . .]"* has been modified to *"[. . .] he then wrote it [. . .]"*. Marginal changes, such as exchanging currencies, insertions/deletions of abbreviations, etc. can be found throughout the news samples and lead to the hypothesis that either a full sentence text is copied or none of it.

6. Conclusion and Future Work

In this paper we presented *News Auditor* , an interactive visual analytics tool that allows the user to compare articles belonging to the same story and understand the differences at three levels of detail. It provides a visual approach to text cluster comparison, interpretation and reasoning. *News Auditor* allows the user to explore text collections as color-coded similarity matrices, which can be controlled by a set of interactions, such as choosing different text similarity measures, sorting and filtering, and expanding rows for a structural comparison. Expanded rows let the user interpret the structural differences, such as article-, paragraph-, and sentence lengths, as well as the textual overlap, represented by a heatmap color-coding of the sentence bars. Finally, a document level view presents two articles side by side to provide full-text comparison showing detail differences by a diff-like view.

News Auditor was applied in two scenarios, using a real world data set. As a part of the future work, we will work on developing more sophisticated methods that will show *why* and *how* was something changed in the report. The tool will be extended to detect and visualize quotations, opinion pieces, and editorial commentaries in the story. We also see the potential for an application in different contexts, e.g. for plagiarism detection, where a text under review could be compared to relevant literature in the field.



AsiaOne-CBSnews-HindustaNEWScomANPRnews-Star-d5fWashingtaljazeeraljazeerbbc-72cacnn-75b1cnn-a448cnn-6f93cnn-95eacnn-316ddailytimdehai-5bearthtimearthtimfinanciaguardianguardianhinduonniafr
22.11 07.12 25.11 07.12 14.12 22.11 07.12 16.12 20.11 07.12 15.12 07.12 07.12 07.12 02.12 19.12 07.12 28.12 07.12 18.11 07.12 16.11 17.11 14.

bbc-72ca 07.12 ⊞ 0.0 ● 0.3 ○ 0.0 ○ 0.4 ○ 0.3 ○ 0.1 ○ 0.3 ○ 0.0 ○ 0.3 ○ 1.0 ● 0.3 ○ 0.3 ○ 0.3 ○ 0.3 ○ 0.3 ○ 0.0 ○ 0.1 ○ 0.4 ○ 0.1 ○ 0.0 ○ 0.0 ○ 0.4 ○ 0.0 ○ 0.3
cnn-316d 02.12 ⊞ 0.1 ○ 0.3 ○ 0.1 ○ 0.3 ○ 0.3 ○ 0.1 ○ 0.3 ○ 0.1 ○ 0.3 ○ 0.3 ○ 1.0 ● 0.4 ● 0.3 ● 0.3 ● 0.3 ● 0.1 ○ 0.1 ○ 0.3 ○ 0.1 ○ 0.0 ○ 0.1 ○ 0.3 ○ 0.1 ○ 0.3
cnn-95ea 07.12 ⊞ 0.1 ○ 0.3 ○ 0.0 ○ 0.3 ○ 0.3 ○ 0.1 ○ 0.3 ○ 0.1 ○ 0.2 ○ 0.3 ○ 0.4 ● 1.0 ● 0.8 ● 0.4 ● 0.3 ● 0.1 ○ 0.1 ○ 0.3 ○ 0.1 ○ 0.0 ○ 0.0 ○ 0.2 ○ 0.1 ○ 0.2
cnn-6f93 07.12 ⊞ 0.1 ○ 0.3 ○ 0.0 ○ 0.3 ○ 0.3 ○ 0.1 ○ 0.3 ○ 0.0 ○ 0.2 ○ 0.3 ○ 0.3 ● 0.8 ● 1.0 … 0.3 ○ 0.0 ○ 0.2
cnn-a448 07.12 ⊞ 0.1 ○ 0.2 ○ 0.1 ○ 0.2 ○ 0.2 ○ 0.1 ○ 0.3 ○ 0.0 ○ 0.2 ○ 0.3 ○ 0.3 ● 0.4 ● 0.4 … 0.2 ○ 0.0 ○ 0.2
cnn-75b1 15.12 ⊞ 0.1 ○ 0.3 ○ 0.0 ○ 0.3 ○ 0.4 ○ 0.1 ○ 0.3 ○ 0.1 ○ 0.4 ○ 0.3 ○ 0.3 ● 0.3 ● 0.3 ● 0.3 ● 1.0 ● 0.0 ○ 0.1 ○ 0.3 ○ 0.1 ○ 0.0 ○ 0.0 ○ 0.3 ○ 0.0 ○ 0.3

Pivot Document: cnn-95ea900303e49aeebf679ac359bc7e73
Compare Document: cnn-6f93ec8577b901ec76ad0c0ca16a20fd
0.8663876711349989

Pivot Document: cnn-95ea900303e49aeebf679ac359bc7e73
Compare Document: cnn-6f93ec8577b901ec76ad0c0ca16a20fd
0.8663876711349989

(a)

cnn-6f93ec8577b901ec76ad0c0ca16a20fd

At the start of the proceedings, Assange was asked for his address and at first gave a post office box.
When told that wasn't sufficient, he wrote a location on a piece of paper and handed it to the judge; it was later revealed that Assange wrote "Parkville, Victoria, Australia" on the paper.
In making his decision to deny bail, the judge cited the fact that Assange gave no permanent address and has a nomadic lifestyle, and that he has access to significant funding that would make it easy for him to abscond.
English socialite Jemima Khan had offered to pay bail of 20,000 pounds ($31,500) and journalist John Pilger also offered a sum of money.
The media was allowed inside the courtroom initially but was later ordered to leave.
Assange appeared in court after turning himself in at a London police station.
He was arrested on a Swedish warrant.
He refused to agree to be extradited to Sweden, so the court now has roughly 21 days to decide whether to order his extradition, said Mark Ellis, executive director of the International Bar Association.
Ellis was surprised that Assange was denied bail, he said.
But Assange will now face an uphill battle to prevent being sent to Sweden --

cnn-95ea900303e49aeebf679ac359bc7e73

At the start of the proceedings, Assange was asked for his address and at first gave a post office box.
When told that wasn't sufficient, he then wrote it on a piece of paper and handed it to the judge; it was later revealed that Assange wrote "Parkville, Victoria, Australia" on the paper.
The judge repeatedly said the case is "not about WikiLeaks," but about serious sexual offenses that allegedly occurred on three separate occasions with two different women.
The media was allowed inside the courtroom initially but was later ordered to leave.
Assange appeared at the City of Westminster Magistrate's Court after turning himself in to a London police station.
He was arrested on a Swedish warrant, though he has not been charged with any crime.
He refused to agree to be extradited to Sweden, so the court now has roughly 21 days to decide whether to order his extradition, said Mark Ellis, executive director of the International Bar Association.
Even though the Swedish warrant is a European arrest warrant designed for easy transfer of suspects among European states, Assange may still fight it. Ellis

Selected Range
Minimum: 11%, Maximum: 100%.

(b)

cnn-6f93ec8577b901ec76ad0c0ca16a20fd

At the start of the proceedings, Assange was asked for his address and at first gave a post office box.
When told that wasn't sufficient, he wrote a location on a piece of paper and handed it to the judge; it was later revealed that Assange wrote "Parkville, Victoria, Australia" on the paper.
In making his decision to deny bail, the judge cited the fact that Assange gave no permanent address and has a nomadic lifestyle, and that he has access to significant funding that would make it easy for him to abscond.
English socialite Jemima Khan had offered to pay bail of 20,000 pounds ($31,500) and journalist John Pilger also offered a sum of money.
The media was allowed inside the courtroom initially but was later ordered to leave.
Assange appeared in court after turning himself in at a London police station.
He was arrested on a Swedish warrant.
He refused to agree to be extradited to Sweden, so the court now has roughly 21 days to decide whether to order his extradition, said Mark Ellis, executive director of the International Bar Association.
Ellis was surprised that Assange was denied bail, he said.

cnn-95ea900303e49aeebf679ac359bc7e73

At the start of the proceedings, Assange was asked for his address and at first gave a post office box.
When told that wasn't sufficient, he then wrote it on a piece of paper and handed it to the judge; it was later revealed that Assange wrote "Parkville, Victoria, Australia" on the paper.
The judge repeatedly said the case is "not about WikiLeaks," but about serious sexual offenses that allegedly occurred on three separate occasions with two different women.
The media was allowed inside the courtroom initially but was later ordered to leave.
Assange appeared at the City of Westminster Magistrate's Court after turning himself in to a London police station.
He was arrested on a Swedish warrant, though he has not been charged with any crime.
He refused to agree to be extradited to Sweden, so the court now has roughly 21 days to decide whether to order his extradition, said Mark Ellis, executive director of the International Bar Association.
Even though the Swedish warrant is a European arrest warrant designed for

Most similiar sentence has matching degree of:
93 Percent.

(c)

Figure 2: *Use Case: Exploration of the same-source content differences. The sorting in the Overview (a) reveals several similar articles published by CNN. The sentence- and word-level differences are shown in (b) and (c), respectively. Detail description of the use case is given in the Section 5.2.*

Acknowledgement

This work was partially funded by the German Research Society (DFG) under grant GK-1042, "Explorative Analysis and Visualization of Large Information Spaces".

References

[Bre12] BREWER C.: Heatmap coloring options - color brewer. http://colorbrewer2.org/, 2012. Online; accessed 25-Feb-2012. 3

[CV07] CILIBRASI R. L., VITANYI P. M. B.: The google similarity distance. *IEEE Trans. on Knowl. and Data Eng. 19* (March 2007), 370–383. 3

[EMM12] EMM: Europe media monitor. http://emm.newsbrief.eu/, 2012. Online; accessed 25-Feb-2012. 2

[ESS92] EICK S. G., STEFFEN J. L., SUMNER JR. E. E.: Seesoft-a tool for visualizing line oriented software statistics. *IEEE Trans. Softw. Eng. 18* (November 1992), 957–968. 2, 3

[Hea95] HEARST M. A.: Tilebars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1995), CHI '95, ACM Press/Addison-Wesley Publishing Co., pp. 59–66. 2

[HM76] HUNT J. W., MCILROY M. D.: *An Algorithm for Differential File Comparison*. Tech. Rep. 41, Bell Laboratories Computing Science, July 1976. 3

[KMS*10] KRSTAJIC M., MANSMANN F., STOFFEL A., ATKINSON M., KEIM D.: Processing online news streams for large-scale semantic analysis. In *DESWEB: 1st International Workshop on Data Engineering meets the Semantic Web* (2010). 3

[KO07] KEIM D. A., OELKE D.: Literature fingerprinting: A new method for visual literary analysis. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)* (2007), n/a, pp. 115–122. peer-reviewed (full). 2

[LBK09] LESKOVEC J., BACKSTROM L., KLEINBERG J.: Meme-tracking and the dynamics of the news cycle. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (2009), ACM, pp. 497–506. 2

[LKS05] LLOYD L., KECHAGIAS D., SKIENA S.: Lydia: A system for large-scale news analysis. In *String Processing and Information Retrieval: 12th International Conference, SPIRE 2005, Buenos Aires, Argentina, November 2-4, 2005: Proceedings* (2005), pp. 161–166. 2

[MJS07] MANKU G. S., JAIN A., SARMA A. D.: Detecting near-duplicates for web crawling. In *WWW '07: Proceedings of the 16th international conference on World Wide Web* (New York, NY, USA, 2007), ACM, pp. 141–150. 2

[OSSK10] OELKE D., SPRETKE D., STOFFEL A., KEIM D. A.: Visual Readability Analysis: How to make your writings easier to read. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST '10)* (2010), pp. 123 – 130. 2

[SFDBC11] SNOWSILL T. M., FYSON N., DE BIE T., CRISTIANINI N.: Refining causality: who copied from whom? In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York, NY, USA, 2011), KDD '11, ACM, pp. 466–474. 2

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on* (1996), IEEE, pp. 336–343. 3

[Wes12] WESKAMP M.: Newsmap. http://newsmap.jp/, 2012. Online; accessed 25-February-2012. 2