Storyfinder: Personalized Knowledge Base Construction and Management by Browsing the Web

Steffen Remus
Universität Hamburg
Hamburg, Germany
remus@informatik.uni-hamburg.de

Manuel Kaufmann
Technische Universität Darmstadt
Darmstadt, Germany
mtk@kisad.de

Kathrin Ballweg
Technische Universität Darmstadt
Darmstadt, Germany
kathrin.ballweg@gris.informatik.tu-darmstadt.de

Tatiana von Landesberger
Technische Universität Darmstadt
Darmstadt, Germany
tatiana.von.landesberger@gris.tu-darmstadt.de

Chris Biemann
Universität Hamburg
Hamburg, Germany
biemann@informatik.uni-hamburg.de

ABSTRACT

This paper presents Storyfinder, an application which consists of a browser plugin and a web server backend with the goal to highlight and manage the information contained in web pages by combining techniques from natural language processing and visual analytics. Webpages are analyzed while visiting them by means of natural language processing components, and metadata in the form of named entities and keywords are extracted and stored for further reference. The extracted information is instantaneously highlighted in the web page and stored in a graph of entities and relations. The graph can be inspected and modified. The investigational scope can be set to a single web page, multiple web pages, or the complete set of analyzed web pages in a user's history. The graph view is designed to adhere to standards of visual analytics and information visualization. Storyfinder is available as an open source application.[1] Its benefit for information access is evaluated in a small user study.

1 INTRODUCTION

The web is nowadays undeniably the major source of a society's information needs. Be it news items, or general facts, the web with its sheer unmeasurable speed of broadcasting new data and its vast quantity of available knowledge is the first choice for information seekers. It is a user's privilege to read or skim a webpage or bookmark it for later reference, but considering that the human memory can be deceptive, it also is a user's obligation to keep information ordered and easily accessible if later reference is required. Instruments exist, such as concept maps[2] [12, 13], or mind maps[3] [4, 5] among many others (see for example [7] for an overview of well-known theoretical knowledge management tools), which provide the necessary methodology and have been implemented in a multitude of prolific, computerized toolkits, which go beyond simple bookmarking.

Another active area of research is *knowledge base induction* from scratch, i.e. plain text documents are processed in order to build knowledge bases. While Navigli et al. [11] or Suchanek et al. [15] (among many others) induce taxonomic or ontological knowledge, i.e. general relations between concepts, [2, 9] follow a more entity-centric approach, i.e. they identify named entities and relations between them and show them in a so-called network of named entities. Kochtchi et al.'s [9] network of names or Yimam et al.'s [17] systems are based on single corpora; as an extension, the system by Benikova et al. [2] produces a network of named entities continuously every day from news texts and incorporates the information from the days before.[4]

With Storyfinder, we aim to support the user to quickly grasp the key concepts of a webpage, make it easily accessible for later usage, and put the new information into relation with previously visited web pages. This strategy helps for comprehending the so-called "bigger picture". Our vision includes supportive investigation of news stories in order to find the links between named entities (NEs) such as persons, locations, organizations or other institutions and facts in form of keywords. From a technical perspective, we address these issues by organizing a user's personally collected knowledge in form of NEs and their relations in the background and present it as a graph while still leaving the option to refine and edit the underlying data for later reference and further investigation. Our approach involves as little user intervention as possible to obtain a network of NEs from a single, currently opened webpage, by analyzing it in the background and providing two visualizations: *a*) highlighting within the webpage, and *b*) entities and their relations are show in a separate graph-based view.

We extend systems like Magpie [6] or ESpotter [18], which only highlight named entities in websites. Whereas Domingue and Dzbor [6] (Magpie) extract named entities by means of semantic web technologies, i.e. the webpage has to encode the entities in the page itself, we follow Zhu et al. [18] (ESpotter) and employ off-the-shelf *natural language processing* (NLP) tools. Due to the archiving functionality, the tool also provides a semantically enriched browsing history.

[1] Storyfinder is released under Apache Software License 2.0 and is available for download at https://uhh-lt.github.io/storyfinder/
[2] http://cmap.ihmc.us/docs/conceptmap.php
[3] http://www.mind-map.com
[4] Network of names: https://ltmaggie.informatik.uni-hamburg.de/non/de, network of the day: http://www.tagesnetzwerk.de/; new/s/leak: http://www.newsleak.io/

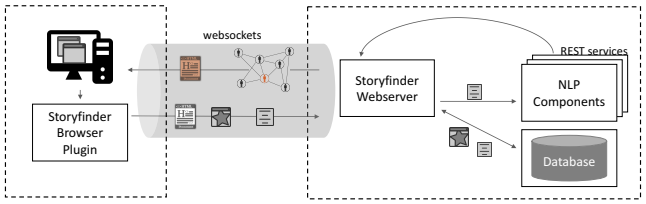


Figure 1: Schema of STORYFINDER's architecture and its components.

2 SYSTEM ARCHITECTURE

Our system consists of three major parts which will be explained in more detail in the next sections:

(1) The **web browser plugin** listens and reacts to user events, initiates the analysis of a webpage, and provides a side pane view with the collected information.
(2) The **server backend** analyzes the webpage, extracts its metadata and stores it for later retrieval.
(3) The **interactive webpage** provides access to the newly gathered information and is embedded in the plugin's side pane view.

The backend is responsible for Information Extraction (IE), whereas the website is responsible for Knowledge Management (KM), and the browser plugin integrates both (IE + KM). A schema of the architecture is illustrated in Figure 1. We describe the components by means of a typical use case: Imagine 'Mary', a web user who is browsing one or more news websites and reads several news articles.

2.1 The Frontend Plugin (IE + KM)

STORYFINDER's browser plugin is responsible for gathering information about the user's current browsing status, i.e. it extracts the html and plain text content of the current webpage, creates a snapshot for archival purposes, highlights the gathered information in an overlay, and provides a side pane in which the interactive webpage (§ 2.3) is rendered in real-time.

After Mary signed in to STORYFINDER and visits the first webpage the plaintext content is extracted using *Readability*[5], a screenshot is created[6], and together with the html content, the data is sent to the STORYFINDER web server, which analyzes and aligns it with the html web page. Once the server has returned the extracted metadata in form of named entities (NEs), keywords, and relations, Mary is able to see the enhanced web page and a graph view of the article in the side pane (c.f. Figure 2).

2.2 The Backend Server (IE)

The server receives the data sent by Mary's browser, processes it and stores the processed data, and, subsequently, returns it to her browser[7]. Three types of information are extracted using standard NLP techniques: 1.) Named entities (NEs), i.e. persons, organizations, locations, and so-called other NEs, 2.) keywords; we henceforth refer to all NEs and keywords as 'entities', 3.) relations between entities. Preprocessing steps such as tokenization, sentence splitting and parts-of-speech (POS) tagging as well as named entity recognition (NER) for English is carried out using the Stanford Core NLP application server [10]. The architecture allows exchanging individual modules, e.g. German texts are processed with GermaNER [3] in favor of better German NE extraction quality. Components are loosely coupled through RESTful web services and easy to deploy due to the usage of Docker[8] containers. Also, the storage engine[9] runs within a Docker container and is thus exchangeable by other, more scalable database implementations if needed.

Figure 2: Screenshot of the default STORYFINDER plugin view. A currently opened webpage is analyzed, the extracted entities are highlighted in an overlay, and rendered in a graph together with their relations in STORYFINDER's interactive webpage, which is shown in a side pane of the browser.

Additionally, keywords are defined to carry a large value of information, hence, we employ a simple, yet effective, keyword extraction mechanism for n-grams up to size three using TF-IDF (term-frequency inverse-document-frequency) measures: Every n-gram with an TF-IDF score above a certain threshold is added to the list of entities with the pre-defined type 'KEYWORD'. The document frequency (DF) is based on all documents seen so far by a certain user, thus, the system is implicitly equipped with a user preference mechanism — or better: a user dis-preference mechanism because a higher DF yields a lower TF-IDF score, the system thus pays more attention to new / unseen information.

For automatic relation extraction (RE) a relation is drawn for every entity or keyword which co-occur in the same sentence. Those relations can then be manually labelled or removed. More sophisticated approaches to RE exist and can be generally integrated by exchanging the RE module.

[5] *Readability* is a Mozilla Firefox functionality for extracting the main content of a webpage removing boilerplate content such as navigational elements, banner, sidebars, header, footer, advertisements, etc.
[6] Using Mozilla Firefox's built-in features

[7] We use *socket.io* in order to maintain a server to client connection.
[8] Docker (http://docker.io) provides convenient deployment of applications by means of operating system virtualization techniques.
[9] We currently use a MySQL database.

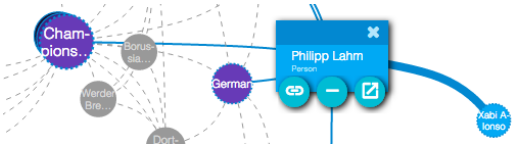


Figure 3: The entity 'Philipp Lahm' is selected, other nodes and edges are grayed out except direct neighboring edges and nodes. Hovering over edges emphasizes them.

Figure 4: Illustration of node labelling heuristics.

2.3 The Frontend Webpage (KM)

Mary's browser receives the extracted information in form of NEs, keywords, and relations and visualizes it in the side pane view showing a responsive web page that is also available directly through the browser. The visualization of entities and their relations in a graph is a crucial aspect in order to reduce the *cognitive load* [16] for a user. Keim et al. [8] explain the *information overload problem*, stating that a user is distracted and overwhelmed by the presentation of irrelevant information for the current task at hand. The entire research field of *visual analytics* (VA) is concerned with the preparation, filtering, and presentation of information in order to overcome this issue. E.g. Ballweg et al. [1] have shown that a graph of NEs visually supports supports the exploration and topical overview of texts and text collections. Storyfinder follows best practices, and adheres to current standards of VA. Some of Storyfinder's key features regarding VA techniques are:

- Different colors for different entity types
- Nodes are visually grouped for multiple occurrences of an entity
- The size of a node represents the relative importance of the entity
- Animations provide immediate user feedback
- Active elements are centered and visually promoted
- Node labels are adjusted to their size; depending on the size of a node we apply different heuristics in order to visualize the entity's name within (cf. Figure 4)

A screenshot of an active node with a selected edge is illustrated in Figure 3. The PageRank algorithm [14] is applied in order to determine the importance of a node. The number of important nodes to show is thresholded by the top n entities depending on the visible area.

In this view, Mary is able to edit the graph, add new nodes either directly through the Storyfinder webpage or by right click in the currently opened webpage. Additionally, Mary is able to navigate to a global graph view, which contains all entities and relations (filtered by PageRank) from all visited webpages, or select a group of web pages and investigate a focused graph. Detailed views of entities and their relations are provided with additional information, e.g. relations are manually labelled with a primary label, which should be generally valid for a relation between two entities, and each occurrence of a relational instance, i.e. the entities occurring in the same sentence, are manually labeled by so-called secondary relation labels (c.f. Figure 5). Furthermore, Mary is supported by a full text search of her history, i.e. she is able to query a string and find all web pages and entities which match the query and navigate to the details of the found article or entity. Since webpages are also stored (screenshot and plain text), Storyfinder serves as a structured browsing history.

3 CASE STUDY

We performed a small case study in which subjects where asked to answer questions about a particular topic which they were unfamiliar of. The questions were unknown beforehand but the general topic has been told, i.e. subjects had the chance to prepare and collect information from the web with the help of different tools. We split the subjects into two groups, one group having access to Storyfinder and one without. We then presented the questionnaire and limited the duration for completion. We then measured the number of correctly completed questions, and since the case study was performed in lab conditions, we were able to log the number of web requests during the session. The subjects in group 1 (w/o Storyfinder) gave 3.5 wrong answers in average, while group 2 (w/ Storyfinder) replied with an error rate of 0.83 in average. Group 1's accuracy is 86.5% while group 2's accuracy is 96.4%, and group 1 had roughly 28% more web requests than group 2. These results clearly indicate the benefit of the Storyfinder tool to grasp the content of webpages and retrieve the personal history.

4 CONCLUSION

We presented Storyfinder, a user based application for information and knowledge management. The tool extracts, highlights, and visualizes NEs and keywords in webpages while browsing the web. The stored information is searchable and editable in an entity-graph centered view which adheres to common standards of VA. By using user profiles, the extracted knowledge is personalized and only privately accessible. PageRank [14] is used as a measure of information and the visualized entities are ranked by a user's preference. While StoryFinder was conceptualized and presented as a browsing tool for news articles, we will extend it for knowledge management in scientific literature in the future where domain-specific article collections can be browsed by methods, citations, datasets or other metadata.

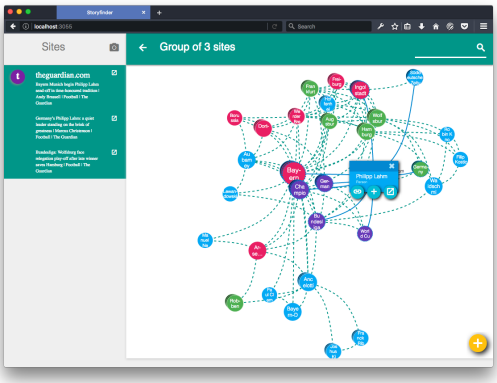
REFERENCES

[1] Kathrin Ballweg, Florian Zouhar, Patrick Wilhelmi-Dworski, Tatiana von Landesberger, Uli Fahrer, Alexander Panchenko, Seid Muhie Yimam, Chris Biemann, Michaela Regneri, and Heiner Ulrich. 2016. new/s/leak – A Tool for Visual Exploration of Large Text Document Collections in the Journalistic Domain. In *VIS 2016 workshop on Visualization in Practice*. 1–14.

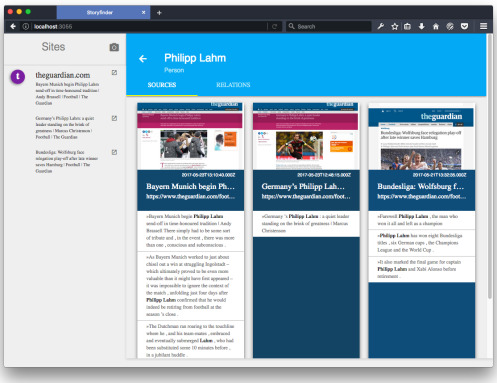
[2] Darina Benikova, Uli Fahrer, Alexander Gabriel, Manuel Kaufmann, Seid Muhie Yimam, Tatiana von Landesberger, and Chris Biemann. 2014. Network of the day: Aggregating and visualizing entity networks from online sources. In *Proc. NLP4CMC Workshop at KONVENS*. Hildesheim, Germany, 48–52.

[3] Darina Benikova, Seid Muhie Yimam, Prabhakaran Santhanam, and Chris Biemann. 2015. GermaNER: Free Open German Named Entity Recognition Tool. In *International Conference of the German Society for Computational Linguistics and Language Technology (GSCL-2015)*. Essen, Germany, 28–31.

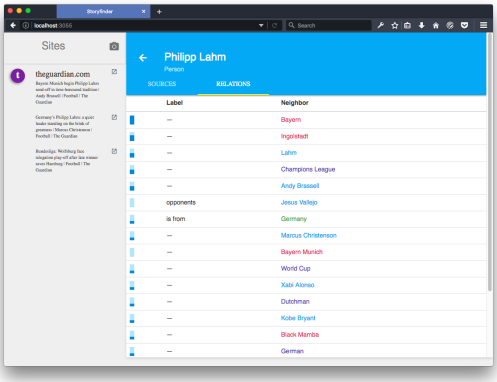
[4] Tony Buzan. 1974. Using both sides of the brain. *Dutton: New York* (1974).



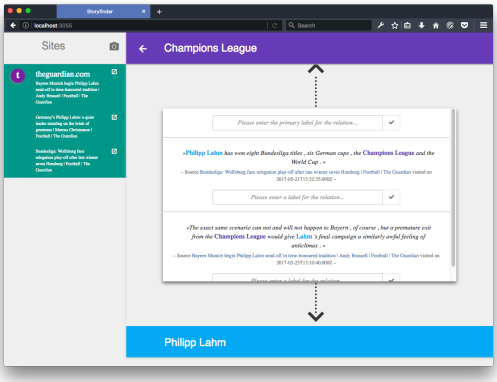
(a) A graph of entities scoping a group of three selected webpages.



(b) Detail view of an entity with its source web page (local copy) and source sentences.



(c) Relations of a selected entity with other entities and their primary relation label.



(d) Enter relation details, i.e. label co-occurrences in one or more sentences. Each relational occurrence can be labeled and a primary label can be specified for generally valid relations between entities.

Figure 5: Detailed views of the collected and managed information.

[5] Tony Buzan and Barry Buzan. 1996. The Mind Map Book: How to Use Radiant Thinking to Maximize Your Brain's Untapped Potential. (1996).

[6] John Domingue and Martin Dzbor. 2004. Magpie: Supporting Browsing and Navigation on the Semantic Web. In *Proceedings of the 9th International Conference on Intelligent User Interfaces (IUI '04)*. 191–197.

[7] Martin J Eppler. 2006. A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing. *Information Visualization* 5 (2006), 202–210.

[8] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. 2008. *Visual Analytics: Definition, Process, and Challenges.* Springer Berlin Heidelberg, Berlin, Heidelberg, 154–175.

[9] Artjom Kochtchi, Tatiana von von Landesberger, and Chris Biemann. 2014. Networks of Names: Visual Exploration and Semi-Automatic Tagging of Social Networks from Newspaper Articles. In *Computer Graphics Forum*, Vol. 33. 211–220.

[10] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Baltimore, Maryland, 55–60.

[11] Roberto Navigli, Paola Velardi, and Aldo Gangemi. 2003. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems* 18, 1 (2003), 22–31.

[12] Joseph D Novak. 1998. *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations.* Lawrence Erlbaum Associates.

[13] Joseph D Novak and D Bob Gowin. 1984. *Learning how to learn.* Cambridge University Press.

[14] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66. Stanford InfoLab.

[15] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web.* Banff, Alberta, Canada, 697–706.

[16] John Sweller, Jeroen JG van Merrienboer, and Fred GWC Paas. 1998. Cognitive Architecture and Instructional Design. *Educational Psychology Review* 10, 3 (1998), 251–296.

[17] Seid Muhie Yimam, Heiner Ulrich, Tatiana von Landesberger, Marcel Rosenbach, Michaela Regneri, Alexander Panchenko, Franziska Lehmann, Uli Fahrer, Chris Biemann, and Kathrin Ballweg. 2016. new/s/leak – Information Extraction and Visualization for Investigative Data Journalists. In *Proceedings of ACL-2016 System Demonstrations.* Berlin, Germany, 163–168.

[18] Jianhan Zhu, Victoria Uren, and Enrico Motta. 2005. ESpotter: Adaptive Named Entity Recognition for Web Browsing. In *Professional Knowledge Management: Third Biennial Conference.* Kaiserslautern, Germany, 518–529.