Erschienen in: Proceedings of LREC 2014 : Ninth International Conference on Language Resources and Evaluation : May 26-31, 2014, Reykjavik, Iceland ; VisLR workshop : visualization as added value in the development, use and evaluation of Language Resources / Nicoletta Calzolari ... (Hrsg.). - European Language Resources Association (ELRA), 2014. - S. 21-28

Enhanced News-reading: Interactive and Visual Integration of Social Media Information

Florian Stoffel, Dominik Jäckle, Daniel A. Keim
University of Konstanz
Department of Computer and Information Science
Box 78
78457 Konstanz, Germany
firstname.lastname@uni-konstanz.de

Abstract

Today, everyone has the possibility to acquire additional information sources as supplement to articles from newspapers or online news. The limitations of classical newspaper articles and restrictions of additional materials on online newsportals often lead to the situation where the reader demands additional news sources and more detailed information. When using the Internet, exploiting new information sources is a trivial task. Besides professionally administered information sources, like for example large newsportals such as cnn.com, there is a growing amount of user generated content available. Services like Twitter, Facebook or Reddit allow free discussion of any subject, giving everyone the possibility to participate. In this paper, we demonstrate an approach that combines professionally generated news content with user-generated data. This approach effectively enriches the information landscape and broadens the context of a given subject. For the presented system, we focus on Reddit, one of the biggest web portals for user-generated contents. Taking the general nature of user generated content into account, we exploit meta-data and apply *Natural Language Processing (NLP)* methods to allow users to filter additional information, which is also supported visually.

Keywords: Challenges in visualizing language data, Visualizations for language resources, Visualization infrastructures / visualization components

1. Introduction

Large parts of today's information gathering processes imply reading professionally written news articles. Depending on the medium and the kind of publication, the articles need to be limited in length and cannot naturally cover the whole context of the article's subject. Nevertheless, they often carry enough information to make readers curious about additional information; for example to be able to assess the background or specific aspects of the subject. Today, these additional information needs can easily be satisfied by visiting different other newsportals or news aggregators like Google News[1] or The Huffington Post[2].

Besides those additional information sources, there is an increasing number of websites available, which allow users to post comments on almost all subjects one can think of. Popular examples are Twitter with more than 200 million active users and over 400 million tweets per day (Wickre, 2013), or Facebook with 728 million daily users (Facebook, 2013). More like a general place for discussions and lesser a social network, Reddit[3] has more than 112 million registered users in January 2014 (Redddit, 2014). These huge numbers of participants lead to large data sets, covering a wide range of topics with different views and opinions. For those data sources, in terms of context and broader information for a specific subject, there is a high chance to find and gather additional information, deviating opinions, and even controversial discussions.

We propose to enrich professional news and articles with contents from such proposed websites. Existing approaches that establish a connection between those different content types (Hu et al., 2012; Gamon et al., 2008; Park et al., 2013; Meij et al., 2012), link news sites with blog posts, tweets, Wikipedia articles, and vice versa. We think, that Twitter is not fully suitable to provide context for given news, mainly because of the short length of posts of maximal 140 characters. Also, the users are using special terms and words to keep their posts short. Posts are in addition changing over time. This adds more complexity to machine-based feature extraction, and requires users, that are not familiar with this type of language, to learn it in order to fully understand a tweet.

To overcome named issues, we decide to use Reddit as one source for social media data. Besides the language issues, this choice also provides us a much richer data source in terms of text length, metadata, and the discourse structure. Apart from having a rating system for each user post, the discourses are already sorted in different so called *Subreddits*. Subreddits are usually dedicated to a specific topic, such as U.S. news, world news, sports, television food or general politics.

We present a user-driven method to retrieve broader context of actual news items. To do so, users can select and combine terms or phrases from news items to form the context of interest. This terms are then used as non-restricted query to retrieve matching Reddit posts whilst the user creates and modifies the context. Our contributions are: (1): A web based framework for linking user-generated social media contents with news, (2): A real-time social media analysis and visualization system, (3): The utilization of contents from Reddit as social media data.

[1] `http://news.google.com/`
[2] `http://huffingtonpost.com/`
[3] `http://www.reddit.com/`

2. Related Work

In the field of social media analysis already exists a huge amount of related research. Because of the vast amount and different types of data, the wide topic range, and a possibly large number of different users who generate data, social media data analysis is also a field with many research disciplines. It ranges from text analysis, development and application of graph algorithms up to text or metadata based visualization. Following, we focus on two fields, which are the most relevant to us. Namely, the classification of social media text, and the visualization of social media data. Besides this social media data oriented literature, we introduce relevant work in the field of linking different sources and data types.

Classification of Social Media Contents. An early work of Pang et al. (Pang et al., 2002) deals with the challenges of analyzing online contents. The authors classify sentiment for online contents using different machine learning techniques. These techniques are evaluated against the simpler method of topic-wise categorization. In (Kim and Hovy, 2004; Pak and Paroubek, 2010) the authors present techniques to automatically create a sentiment classifier based on human labeled test data and posts from the microblog plattform Twitter. A different goal is pursued by Yang et al. (Yang et al., 2007). In this approach, the authors perform emotion classification with the help of support vector machines and conditional random fields on blog data. The authors found out, that the last sentence describes a posts emotion best.

All these techniques concentrate on the classification of documents from online or social media sites using state-of-the-art machine learning techniques. A simpler approach is the usage of affective word lists as presented by Nielsen (2011). The author also elaborates on the differences between results which use his word list and an elaborated machine learning techniques from Thelwall et al. (2010). Surprisingly, the word list performs almost on par with machine learning techniques. An extensive report on state-of-the art technologies can be found in (Pang and Lee, 2008).

Since we want to build an interactive system that incorporates the classification of social media data, we decided to use the text classification presented by Nielsen (2011).

Visualization and Social Media Data. Hao et al. present a combination of sentiment analysis and social media contents in a streaming environment (Hao et al., 2011). The authors apply topic-based sentiment analysis to streaming data, which is used to generate two kinds of Twitter microblog visualization: a so called *Sentiment Calendar* and a *Sentiment Geo Map*. These two visualizations display data with temporal or geographical context.

A compact visualization of opinion analysis outcomes of user-contributed text data is presented by Oelke et al. (Oelke et al., 2009). To do so, the authors create a pipeline to extract application specific features, which are visualized compact to be able to provide reports out of the generated data. The combination of different techniques for sentiment analysis and the visualization of their outcome is the subject of Gamon et al. (Gamon et al., 2005). Starting with an application-dependent taxonomy, the authors describe the combination of classification, clustering, and a Tree Map visualization as part of an interactive system. Similar analysis goals are pursued in (Abbasi and Chen, 2007), where a system, capable of analysis of user-generated text data from different online sources, is shown. Having defined different features on three levels, namely style, topic and sentiment, the visualization is generated on text level and plotted into the document background.

Besides visualizing text features extracted from social media data, it is also possible to create visualizations for non-textual features. Smith et al. (Smith et al., 2009) present a system which visualizes the user base of a social media network with graph layouts and algorithms from graph research.

In excerpt from relevant literature, it becomes clear, that the actual visualization is application-dependent. If the task is to provide detailed insights into extracted features, a detailed visualization like the Ink Blots (Abbasi and Chen, 2007) is appropiate. Since we are aiming at overview or report style visualization, our approach is more like the work presented by Oelke et al. (Oelke et al., 2009). A high level overview of visual analysis of social media data can be found in (Schreck and Keim, 2013).

Linking News and Social Media. An investigation whether there exists a link between social media, in this case Twitter, and real world news is presented by Hu et al. (2012). The authors come to the conclusion that this is the case, and there can be even different types of users with respect to news spreading for the observed case. A tailored technique for merging political news with blog posts is proposed by Gamon et al. (2008). This technique can be described as supervised, as some domain knowledge is incorporated before linking takes place. Park et al. show a case study for political elections, where a link between real world news and Twitter messages is used in order to explore social media data regarding elections (Park et al., 2013). To link tweets and Wikipedia articles is the goal of the work of Meij et al. (2012). The authors extract concepts and different text and metadata features, and use a machine learning approach to establish a link between those different data sources. In (Tsagkias et al., 2011), the authors introduce a methodology to link news contents with social media contents in general, which also serves as the foundation for our work.

3. System and Process Pipeline

Today's news websites provide a huge amount of news articles and data, but they do not focus on the user's information needs. This can result in long lasting news browsing sessions to get additional insights or new information in general, which may also lead to sites outside the starting news source. In contrast, we present a user-centered system providing a link between social media data and a user-defined context generated from news articles. In this chapter, we motivate the approach and describe our process which is modeled along a pipeline.

Data Science and Visualization have already been brought to the web and are therefore widely spread. The creators

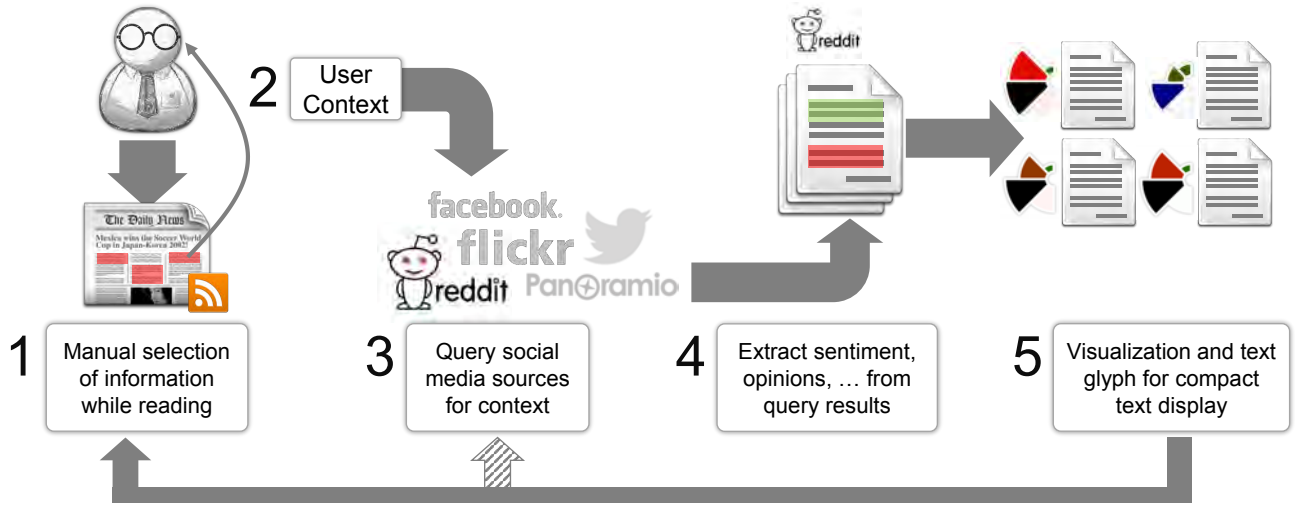

Figure 1: Five step pipeline. (1) The user reads different sets of news articles, possibly from different sources. To (2) create a context, the user either uses the integrated search functionality or manually selects words or phrases with the mouse. (3) Based on the selection, social media data providers are linked to right into the system. (4) Once these resources have been queried, we apply text analysis to extract scores from the retrieved data. (5) The retrieved data is visualized with a compact glyph, and some additional information is shown. After (5), the user can decide to go back to news reading, or to create links to other social media data providers.

and innovators in that field enjoy great reputation for their work, but today's web-based data visualizations are either static or contain interaction techniques that do not suit the user-driven, visualization supported analysis which is called Visual Analytics. This is a challenge for our framework, both, for the visualization and the text processing side. The web-, and as a consequence also the browser-based implementation of the framework has been chosen, because it opens possibilities of portability and accessibility not possible with locally running, native programs. There is also the advantage of being able to deploy the analysis infrastructure at a central location outside of the actual web interface, which makes it easy to get the system running on a smart phone or tablet device.

3.1. Process Pipeline

The framework follows the process pipeline shown in Figure 1, which defines the user interaction and the linking of a user-generated context with social media data sources. In the beginning, the user selects relevant news sources which are merged and displayed together (1). In order to create a context describing further information needs of the user, parts of the displayed news can be selected with the mouse (2). Besides that, a free text input is also available. In the next part of the pipeline, the link between the user-generated context and social media data providers is established (3). To do so, a provider specific query is generated and executed. After the query results have been retrieved, the data is analyzed and different scores are computed for each of the returned pieces of data (4). In order to provide a responsive interface, these analysis steps have to meet strict processing time and resource requirements. The extracted scores are visualized as glyph besides a compact display of the retrieved data (5). At this stage of the pipeline, there are two possibilities to continue. The first is to examine the retrieved data, which refers back to (1). To enhance this task, various filters, orderings and an adjustable scoring are available. Another possibility is to introduce a link between different social media data providers to broaden the retrieved results. Currently, this possibility is part of our pipeline, but not implemented in the framework.

3.2. Integration of Multiple Data Sources

There are two stages in the pipeline (Figure 1), which rely on the combination of multiple data sources. The first one happens in (1), where the content of user-defined news sources is merged in one display (Figure 2, (2)). The second part, where multiple data sources are integrated, is indicated by the feedback from (5) to (3), the shaded arrow in Figure 1. This is, where the linking from one external data source, i.e. not a news source, to another external social media source should take place. The combination of data sources requires the ability to extract and retrieve the same information and metadata, like title, text, date, tags and more. This is one of the biggest challenges, since those information extractors are specific to the each of the data providers. Most of them use similar techniques for retrieving the data (JSON-RPC, SOAP, or REST), but the returned data format can differ heavily.

3.3. User-Driven Knowledge Discovery

Users play a crucial role in all kinds of Visual Analytics applications. Therefore, it is worth to explicitly mention the steps of the pipeline in Figure 1 involving the user again. The user has to define the news feeds which should be presented (1). Additionally, a context to use for linking social media data sources has to be defined manually (2). The visualization had to be interpreted (5), and there is the optional filtering and ordering, which is completely user-driven (5) $\rightarrow$ (1), (5) $\rightarrow$ (3). By involving the user at those crucial steps of the pipeline, it is ensured that the results are

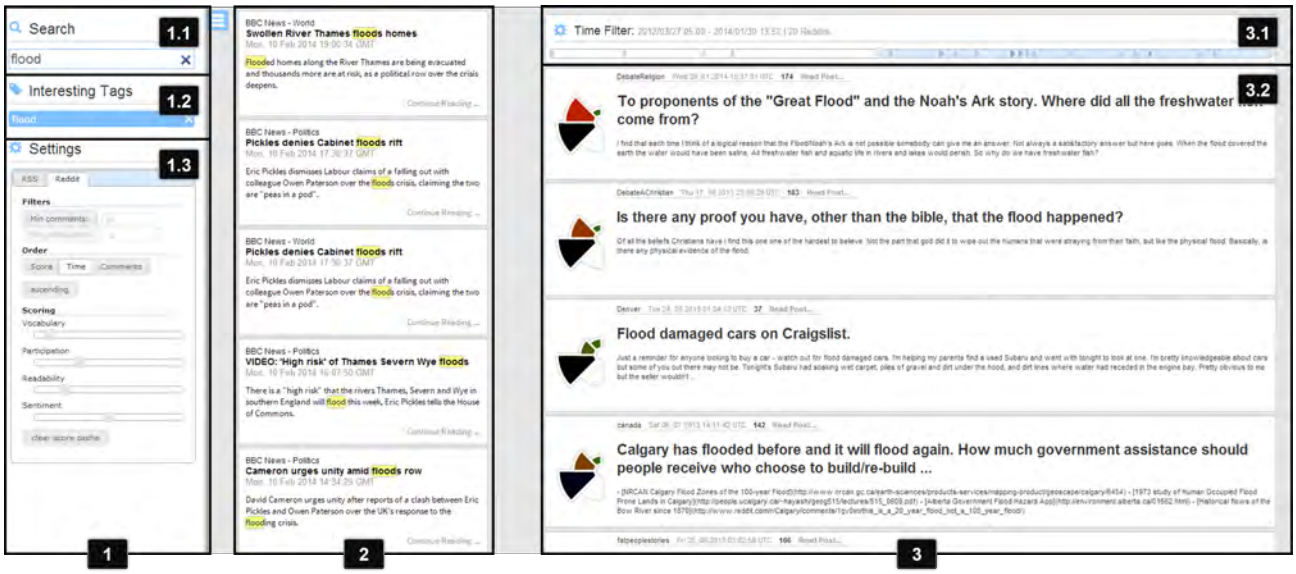

Figure 2: Our system consists of three main parts, placed next to each other from left to right. (1) System interactions: (1.1) Search for relevant terms within all displayed news articles, (1.2) manage tags used to create the context the user is interested in, and (1.3) allows further customization of the display in (3), for example the ordering or apply in-place filters. (2) The news data is presented in descending order over time. It changes automatically when the news sources are changed by the user. (3) Following the user's context, data from social media providers are presented. These can also be sorted by the user either (3.1) by time or (1.3) with the help of various ranking and filters.

tailored to the user's interests, and not by parameters from algorithms which are hard to understand.

3.4. System Design

Figure 2 gives an overview of the system implementing the pipeline. It is built out of three main elements: a sidebar offering various interaction and configuration possibilities (1), the news article display (2), and the augmented social media display (3). When the application is started, the user can adjust the active sources in (1.3), which are then merged in the combined news display (2). This list can be browsed through and used to build the context by selecting words or phrases with the mouse, which is displayed in (1.2). Additionally, there is a full text search feature available, which interactively filters the displayed news items (1.1). Every time the context is changed, the link to the social media data sources, currently only Reddit, is established and the results are retrieved asynchronously and displayed in (3), augmented with the glyph. The display can be adjusted by the user in terms of the order, and some in place filtering can be applied (1.3).

3.5. Text Analysis and Scores

To be able to judge whether a returned post is worth reading or of general interest, a text analysis is performed. The analysis has to meet strict resource and runtime limits. Otherwise, the system would slow down the process of knowledge discovery, which results in not fulfilling the intended task or a displeasing slow down in the interaction. Currently, these limits are met because the text analysis methods are not implemented by using any resource or computation intensive technologies. To compute text readability, the well established and efficiently computable Flesch Reading Ease measure is used (Kincaid et al., 1975). A measure for the complexity of the used vocabulary is introduced, to take into account not only the quantitative, but also the qualitative aspect of the text To do so, the share of words not contained in the 10,000 most frequently used English words, extracted from the Project Gutenberg corpus[4], is computed. To provide a sentiment score, the positive/negative wordlists from Hu et al. are used (Hu and Liu, 2004). The sentiment score is computed by subtracting the number of negative sentiment words from the number of positive sentiment words. To express the comprehensiveness of the discourse, which we assume is positively correlated with the number of participants in the discussion, the average number of posts per user is computed. Each of the described features is computed on sentence (readability, sentiment) or post level (non-basic vocabulary, number of posts per user) respectively. Those features have been chosen, because they give an insight in the quality of the text and discourse, which in our point of view are relevant when users decide if a text is worth reading or not.

3.6. Visualization of Analysis Results

The visualization for the text analysis output is shown in Figure 3. Each text feature is represented in a separate sector, which are equally sized when they are at their maximal extent. The feature value is double encoded in the size of each segment and the color value. For each of the segments, the feature value and color is modified separately to optimize its display. For example, the sentiment is mapped to colors ranging from red (negative) to white (neutral) to green (positive). A segment displaying a neutral sentiment score is covering roughly one fifth of the available, maximal segment area. If the score is negative or positive, the segment is

[4] http://www.gutenberg.org/

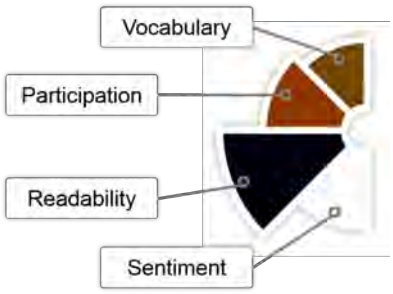

Figure 3: The glyph used to display the text analysis results, which are computed for each data item separately.

enlarged according to reach its maximal size when the text is classified as maximal negative or positive. Since different colors are assigned positive and negative sentiment, the actual sentiment is easy to determine. The readability score is mapped to a color scale from white (easy to read), blue (medium readability), black (hard to read). We chose non signal colors to not induce any ranking like a text is really hard to read, when the segment is colored red, since the actual readability is subjective. For the level of used words and the participation score, we use a color map ranging from green to red, because in those cases we want to support the user in his selection of text to read.

4. Application Example

This application example illustrates the process of broadening a given context and generating new possible exploration directions of a given subject.

At first, the user begins to browse a news stream retrieved from ESPN, a popular sports news site. In a news item, the term *Olympic* sticks out and stirs up the user's curiosity. To add this term to the current user context, the term can be selected with the mouse (Figure 4 (1)). Immediately after changing the current user context, the social media sources, in this case Reddit, is linked and posts regarding the context; in our case, the term *Olympic* is retrieved. Since only recent posts are of note, the time filter is applied to the results by only selecting results that begin with the middle of October 2013 (Figure 4 (4)). Right below, the generated overview from the social media data, enriched with the glyph, is presented. Additionally, the order of the displayed social media data is adjusted by the user-defined scoring, where the high user interaction is rated highest, followed by the sentiment, readability and vocabulary score (Figure 4, (3)). In the first post, the name *Sochi* sticks out, which attracts the users further interest.

To continue with the new context addition *Sochi*, the user applies the integrated full text search, Figure 5 (1), to filter the news items accordingly. Also, the term is added to the user context, Figure 5 (2). This causes an immediate update of the displayed social media posts, as presented in Figure 5 on the right hand side.

In the news, which are now filtered by the term *Sochi*, a shift of providers can be seen. The news displayed in Figure 4, originated from ESPN, as the user has decided to read sports news. In Figure 5, the displayed news item originate from CNN only. As it can be seen on the example Figure 5 (3), there new perspectives arise from the new user context, shifting from sports to politics.

5. Discussion

Currently, the system is designed to support one task: the extension or generation of social media context given a set of terms, manually selected by the user from news items. Not taking the whole news item, but only a set of user selected terms for context generation is the guarantee, that the system only retrieves contents which the user is interested in. Having that terms and the link to a social media data provider, one can think of additional tasks besides the context extension.

A number of other tasks can be supported by the system and the underlying technologies. When trying to describe them, four categories have to be considered separately: At first, users must be able to formulate the context they are interested in. Currently, there is the possibility to select terms that describe the context by clicking and dragging on the news item text. To allow terms which are not contained in the news item, but relevant for the context according to the user, a manual term input is available. This works reasonably well for describing the desired context with text, but for other data sources that do not provide text data other input methods must be available. The second category describes the technology used for linking a social media data source, since it must be possible to establish, query and retrieve the data type provided by the source. In our case, we use the search api from Reddit, which returns different types of data for one query in the same JSON based data format. The third category covers available interactions. It is obvious, that for exploration tasks, the system should be able to establish not only a link between user-defined context terms, but also between different social media sources, as it has been already formulated in the pipeline on Figure 1. This should be covered by the available interaction techniques and also the type of user interface. In the current state, we display the results from the social media provider as a list with a glyph describing each result visually. This list can not be used for further exploration of the context, because it doesn't allow direct selection of terms from the results. For an exploration task, this should be possible. The fourth category covers the user interface. To stick with the exploration task example, having the possibility of establish a link between different social media data sources, it should be possible to use that link in both directions, if a dead end of the exploration has been reached. A simple solution for linking in both directions is to extend the user interface by providing both ends of the link. In the current system, this would mean that two lists with postings are shown next to each other, and both provide the same interaction techniques.

6. Conclusion and Future Work

In this paper, we present a framework to link news and social media content. Our application example shows this combination for newsfeeds from ESPN and CNN, combined with the social media platform Reddit. We allow users to define the context they want to explore by themselves, thus not imposing any direction or things they are not interested

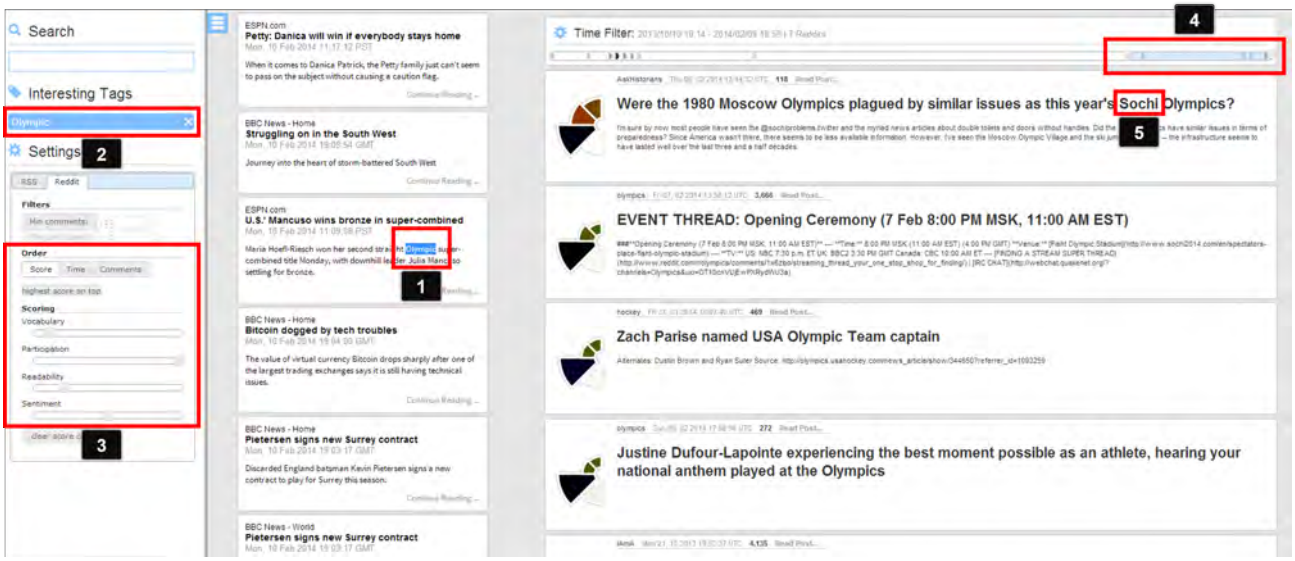

Figure 4: First step in the process towards enhanced news reading. When starting to use our application, the user is presented a temporal ordered news stream that consists of multiple news sources. ESPN sports-news-stream publishes articles about the Olympic games that take place in 2014. The user is interested in additional information and therefore (1) manually highlights the term *Olympic* within the article. The selected term is automatically transfered and added to the (2) list of interesting tags. Meanwhile, social media content matching the given keyword is retrieved. To rank the social media results according to the user's preferences, the user changes the (3) scoring order and applies the (4) time filter. The top ranked result reveals the (5) venue of the Olympics: *Sochi*.

in, but can be rated high when automatic methods decide which content to link.

We chose Reddit, because it provides much richer content than other social media data sources like Twitter. Also, accessing data is not restricted in any ways, at is can be on platforms like Facebook. It also has the advantage to be actively moderated, so the resulting content should be free of any off-topic contents, troll postings, or similar internet phenomena.

Compared to existing work, we bring the concept of linking news and social media directly to users. In our framework, we apply it to news reading in form of providing a broader context and information, which is usually not included in articles from news sites. Especially, when the integration of further data sources, different types and the linking among them has been added, the additional information provided by our framework must not only be presented as text, as it is the case now. For example it is then possible to represent additional context information as image tiles, which are associated with the corresponding social media contents. This would reduce the cognitive load required for users to actually generate new knowledge about the subject they are interested in.

6.1. Future Work

In the future, we want to expand the capabilities of the system to generate a broader social media context. To do so, there is the possibility of using the user-provided context as seed for a classical query expansion. Sources like WordNet or DBPedia can be used to expand the context before actual queries are created and sent to the social media data providers.

We are also interested in the possibility to add other data types than text. Currently, we have a compact, row like representation of social media text data, which is displayed with a title on the right, see Figure 2, 3.2. There are other social media data sources available, like Flickr, Panoramio, or Instagram, where annotated image data is available. Also, with the images, one could think of a dashboard like representation of the social media data, represented by associated images. In terms of the actuality of the retrieved data, it is without doubt that Twitter would also be an optional, valuable addition to the data sources.

The basic text processing, which has been chosen because of the desired real-time capabilities of the linking, could be extended by introducing a two staged analysis model. The first stage represents our current processing, capable of computing results almost instantly after the data had been retrieved. More elaborated methods, like the combination of machine learning techniques for identifying positive and negative arguments, or to analyze the type and structure of the discourse, act as second stage. This additional information can be used as another dimension in the glyph visualization, the data filtering capabilities, or to determine the order of the retrieved data in a fine grained way which is currently not possible.

It is also an open question, whether we can apply the same measurements and methods to different social media data sources. The first problem which can arise, is the type of language and the length of posts given two sources like Twitter and Reddit. An automatic evaluation of different methods and their result should make it clear, whether we need to employ different analysis techniques for different data sources.

The glyph that represents the classification score of each social media entry, helps users to rapidly identify valuable

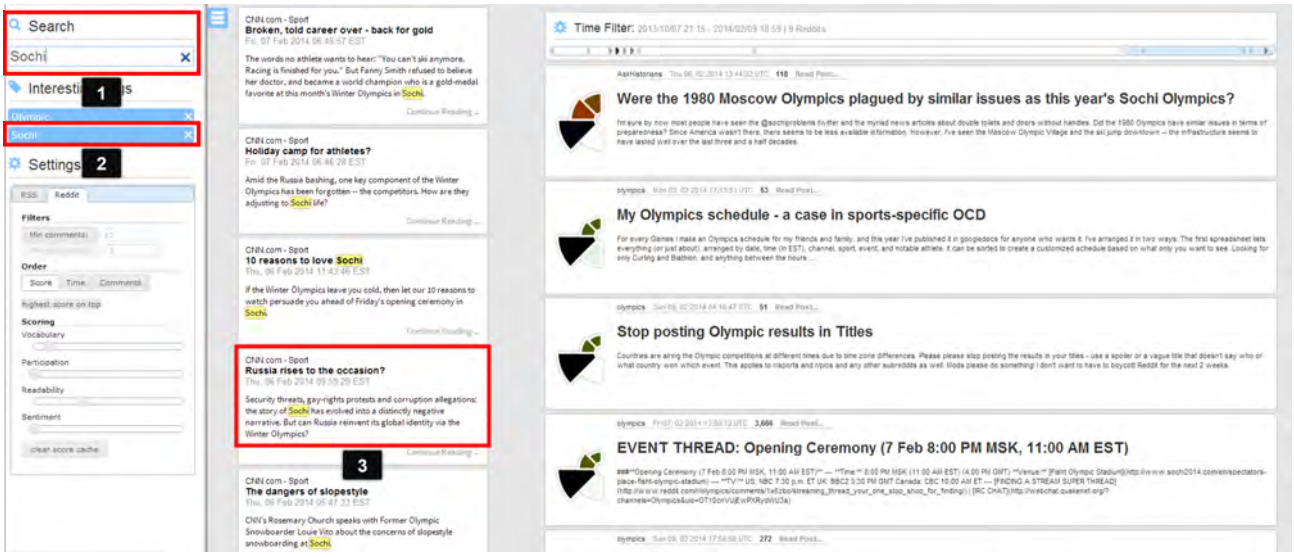

Figure 5: In the second step towards enhanced news reading, the user (1) searches the news for the keyword *Sochi* (determined in Figure 4). In accordance with the query, the news stream is filtered and presents articles that contain given keyword only. Additionally, the shortcut *shift+return* adds the keyword to (2) the list of interesting tags and refines all social media results. The filtered news stream reveals an (3) interesting article published by CNN, named: Russia rises to the occasion? An article about security threats, gay-rights, etc. which gives dissimilar impressions. That finding serves as entry point for further enhanced exploration.

entries, but it lacks of overview. It is obviously hard to visually group entries, especially with more than four categories. Hence, future work includes enhancements of the proposed glyph.

A formal task definition, the required components, analysis techniques and interaction possibilities can be a step towards making our technique more applicable in general. Currently, the supported task is to retrieve context information for a given news item. But given the underlying technology, it is worth exploring different possibilities when the link between a news item and social media data providers has been established. One possible task that builds on the current basis, is the generalization of the system for a broader investigation of a given subject. For this task, extensions of the supported data sources and the interactions, for example to effectively support browsing through retrieved data, need to be added. Nat at last, we want to conduct a controlled user study to see whether the tasks supported by our system are useful for real users.

Acknowledgments

This work has been supported by the German Research Foundation (DFG) project “Feature-based Visualization and Analysis of Natural Language Documents” (VisADoc).

7. References

Finn Å. 2011. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In M. Rowe, M. Stankovic, A. Dadzie, and M. Hardey, editors, *#MSM*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98. CEUR-WS.org.

A. Abbasi and H. Chen. 2007. Categorization and Analysis of Text in Computer Mediated Communication Archives Using Visualization. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL ’07, pages 11–18, New York, NY, USA. ACM.

Facebook. 2013. Facebook Reports Fourth Quarter and Full Year 2013 Results. http://investor.fb.com/releasedetail.cfm?ReleaseID=821954. Online Report, accessed: 2014-02-08.

M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. 2005. Pulse: Mining Customer Opinions from Free Text. In *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis*, IDA’05, pages 121–132, Berlin, Heidelberg. Springer-Verlag.

M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. König. 2008. BLEWS: Using Blogs to Provide Context for News Articles. In E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, editors, *ICWSM*. The AAAI Press.

M. C. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim, L. Haug, and Mei-Chun Hsu. 2011. Visual sentiment analysis on twitter data streams. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 277–278, Oct.

M. Hu and B. Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’04, pages 168–177, New York, NY, USA. ACM.

M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma. 2012. Breaking News on Twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’12, pages 2751–2754, New York, NY, USA. ACM.

S. Kim and E. Hovy. 2004. Determining the Sentiment of Opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING ’04, Stroudsburg, PA, USA. Association for Computational

Linguistics.

J. P. Kincaid, R. P Fishburne Jr, R. L. Rogers, and B. S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, DTIC Document.

E. Meij, W. Weerkamp, and M. de Rijke. 2012. Adding Semantics to Microblog Posts. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 563–572, New York, NY, USA. ACM.

D. Oelke, M. C. Hao, C. Rohrdantz, D. A. Keim, U. Dayal, L. Haug, and H. Janetzko. 2009. Visual opinion analysis of customer feedback data. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 187–194, Oct.

A. Pak and P. Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

B. Pang and L. Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Souneil Park, Minsam Ko, Jaeung Lee, Aram Choi, and Junehwa Song. 2013. Challenges and Opportunities of Local Journalism: A Case Study of the 2012 Korean General Election. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 286–295, New York, NY, USA. ACM.

Redddit. 2014. About Reddit. http://www.reddit.com/about/. Online Web Page, accessed: 2014-02-08.

T. Schreck and D. A. Keim. 2013. Visual Analysis of Social Media Data. *Computer*, 46(5):68–75, May.

M. A.. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, and E. Gleave. 2009. Analyzing (Social Media) Networks with NodeXL. In *Proceedings of the Fourth International Conference on Communities and Technologies*, pages 255–264, New York, NY, USA. ACM.

M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. 2010. Sentiment in Short Strength Detection Informal Text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558, December.

M. Tsagkias, M. de Rijke, and W. Weerkamp. 2011. Linking Online News and Social Media. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 565–574, New York, NY, USA. ACM.

K. Wickre. 2013. Celebrating #Twitter7. https://blog.twitter.com/2013/celebrating-twitter7. Online Blog Post, accessed: 2014-02-08.

C. Yang, K. Lin, and H. Chen. 2007. Emotion Classification Using Web Blog Corpora. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 275–278, Washington, DC, USA. IEEE Computer Society.