(i)DiversiNews – A STREAM-BASED, ON-LINE SERVICE FOR DIVERSIFIED NEWS

Mitja Trampuš, Flavio Fuart, Jan Berčič, Delia Rusu, Luka Stopar, Tadej Štajner
Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
Tel: ++ 386 1 477 3528; fax: +386 (01) 477 38 51
e-mail: {name.surname}@ijs.si

ABSTRACT

With the ever-increasing ease and speed of opinion exchange, the internet often displays the echo chamber effect. This is exacerbated by a free market: search engines and other data aggregators are monetarily incentivized to primarily show the most popular opinions. We propose a data aggregation, processing and retrieval system to combat this phenomenon in the domain of web news. We developed two applications (web, iOS) that allow users to explore news articles along several uncommon dimensions, diversifying them and discovering new aspects of a story. The iOS application works in real time, making for a novel alternative to classic news reader apps. Our user study shows a need for such diversity-aware approaches and judges our solution to be directly useful.

1 INTRODUCTION

The RENDER European project[1] aims at developing tools and services which enable the analysis of text in a Web-based environment from a diversified perspective. The focus is on two main sources of information diversity. Firstly, we are interested in analysing *information content diversity*, and identifying the main topics of the text, its geographical provenance, the opinions expressed in text, and aggregating this content in a short summary. Secondly, we want to observe *diversity in information usage*, and identify how different user groups such as domain experts, user communities or the general public are interacting with RENDER technology.

As a step towards reaching these goals, we propose a case study which is based on near real-time analysis of news. We developed a web application[2] and an iOS client that allow users to browse and summarize news from different perspectives. The individual news articles are grouped into several news clusters. Given a certain topic of interest, the tool provides a succinct summary of the most related news cluster, as well as the individual news articles that have been summarized, sorted by relevance. The user can further specify which perspective on the news should be emphasized: articles containing certain predominant keywords, articles that belong to a certain geographical region, or articles with a positive or negative outlook.

Related work. There are several applications which aim at representing information from different perspectives. *DisputeFinder* [1][2] is a browser extension that alerts the user when detecting that the information accessed is disputed by a trusted source. The tool highlights known disputed claims and presents a list of articles that support a different point of view. *Social Mention*[3] is a social media search and analysis platform which aggregates different user generated content, providing it as a single information stream. The platform provides sentiment (positive, negative, and neutral), top keywords, top users or hashtags related to the aggregated content. The *Global Twitter Heartbeat*[4] project performs real-time Twitter stream processing, taking into account 10% of the Twitter feed. The text of each tweet is analysed in order to assign its location. A heat map infographic displays the tweet location, intensity and tone. *Europe Media Monitor* [3] represents a number of news aggregation and analysis tools that track stories across time, languages and geographic locations. It also detects breaking news stories and hottest news topics. Topic-specific processing is used, for example, to monitor EU policy areas[5] and possible disease outbreaks[6][4].

The remainder of this paper showcases our applications and is structured as follows: in section 2, we describe the data collection and preprocessing shared by both applications. Section 3.1 describes *DiversiNews*, and section 3.2 describes how it was extended *iDiversiNews*, a mobile iOS application, to show near real-time news. Section 4 shows the results of a UI study.

[1] http://render-project.eu/
[2] http://aidemo.ijs.si/diversinews/
[3] http://www.socialmention.com
[4] http://www.sgi.com/go/twitter/
[5] http://emm.newsbrief.eu/
[6] http://medisys.newsbrief.eu

2 DATA COLLECTION AND PREPROCESSING

The data is collected using the JSI Newsfeed. The system's reference article [5] describes how the data sources, mainly RSS, are collected and crawled. It also details some of the preprocessing built into the newsfeed, notably cleartexting and language detection. However, for the purposes of (i)DiversiNews, news stages of preprocessing have been added that have not yet been documented.

Publisher geolocation. We try to associate each publisher/site with geographic coordinates. We crawl public listsings of news publishers to learn the city and country of origin. Failing that, we have developed a set of heuristics that query a WHOIS server with the publisher's hostname and extract the most likely country of origin. Hostnames with national TLDs are automatically assigned to that country. A publisher with a known country but unknown city is mapped to the geocenter of the country.

Stable enrichment. Articles are enriched using the Enrycher [6] service as pointed out in [5]. In the scope of (i)DiversiNews development, several critical stability bugs were fixed and a separate instance of Enrycher was installed for the needs of the project. It performs part of speech (POS) tagging, sentiment analysis, named entity extraction and resolution and DMOZ classification.

Sentiment analysis. The sentiment analysis module in Enrycher has been rewritten from scratch, largely motivated by (i)DiversiNews. A supervised method is now being used with significantly improved performance.

Article clustering. A stream clustering method has been developed by Janez Brank for clustering news articles into stories. It maintains the centroids (in the high-dimensional bag-of-words space) of several thousand clusters using a dynamic proximity search data structure. Each new article is assigned to the cluster with the nearest centroid. If the resulting cluster is large enough, it is periodically considered for splitting into two subclusters using bisecting k-means; the decision on whether to accept the split or not is based on a Bayesian information criterion). Individual articles' weight/contribution to the centroid is attenuated exponentially to prevent old stories from lingering in the system for too long. Overly old articles are discarded. Periodically, we examine pairs of similar clusters and consider merging them. A combination of cosine distance and Lughofer's ellipsoid-overlap criterion is used to determine whether to perform the merge. The service has a push API to keep subscribers updated about cluster membership changes.

We limit ourselves to English articles, although the only language-dependent component is sentiment analysis so expansion to other languages is feasible.

Architecturally, the enrichment is performed in the scope of Newsfeed. Its output is the starting point for DiversiNews and iDiversiNews. The former being designed for browsing through historical data and the latter serving real-time data, their respective caching mechanisms and backends are different, as are obviously the frontends.

3 USER INTERFACES

News data offers many aspects of diversity and no single application can present them all due to sheer information overload. We therefore had to choose only a few and did so based on several criteria: 1) how well defined the aspect is, 2) how good are the automated methods at extracting it and 3) our ability to propose an intuitive user interface for navigating the space of that aspect. In the end, we chose a) topic of focus, b) geography of publisher origin and c) sentiment.

We developed two applications that allow a user to navigate

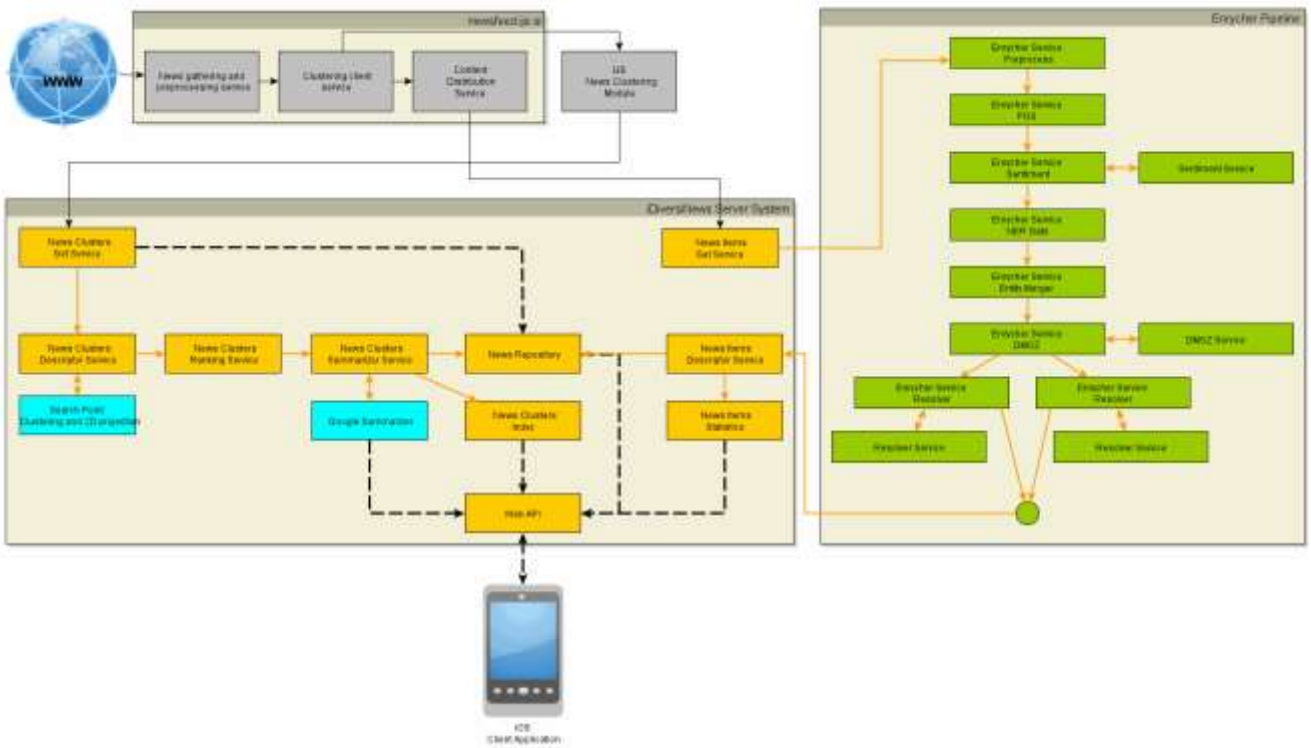

Figure 1: iDiversiNews System Architecture, in part shared with DiversiNews

a cluster of articles along these three dimensions. For example, browsing a cluster of news about the British royal wedding, the user might express interest in articles a) focusing on Kate's headwear b) with a positive spin and c) coming from Japan; the user interface will react by displaying the relevant articles and their summary.

The most prominent part of both applications is the News Story Exploration Area where different views on a same story can be explored. Users select a position in a multi-dimensional story space: by selecting most descriptive keywords ("Search Point" area), pinpointing a particular geographic area (publisher location) and selecting news sentiment (a scale that goes from negative through neutral to positive). The selected viewpoint is then used to rank news articles by relevance. A summary is created from the top scoring articles. Figures 2 and 3 show a sample view of the user interface in DiversiNews and iDiversiNews, respectively.

3.1 DiversiNews (webapp)

This application allows browsing of static news data dumps in a web browser. The browser is a thin client; the other component of the application is a server that fetches data from the Newsfeed and then performs data storage, retrieval, ranking, summarization and result caching. Storage and retrieval are implemented using the QMiner in-memory database engine developed by Blaz Fortuna. Ranking, result caching (= retrieved, ranked articles and their summary) and client-server communication are implemented within the SearchPoint framework. Two summarization algorithms were implemented in the scope of RENDER and are available in the interface; the first [7] was developed by Google and is based on a probabilistic model, the second by JSI and is based on an iterative greedy approach operating in the space of subject-verb-object triplets, using WordNet as background knowledge to define distances between triplets.

Figure 2 shows the user interface. Diversity controls can be found on the right (top to bottom: topic, geography, sentiment); the ranked articles and the summary appear on the left. The entry field on top provides keyword-based search (for example, search for New York, then use diversity controls to discover articles about NY coming from different parts of the world with different focuses and sentiments). It is also possible to analyse all articles belonging to the same story cluster (not shown).

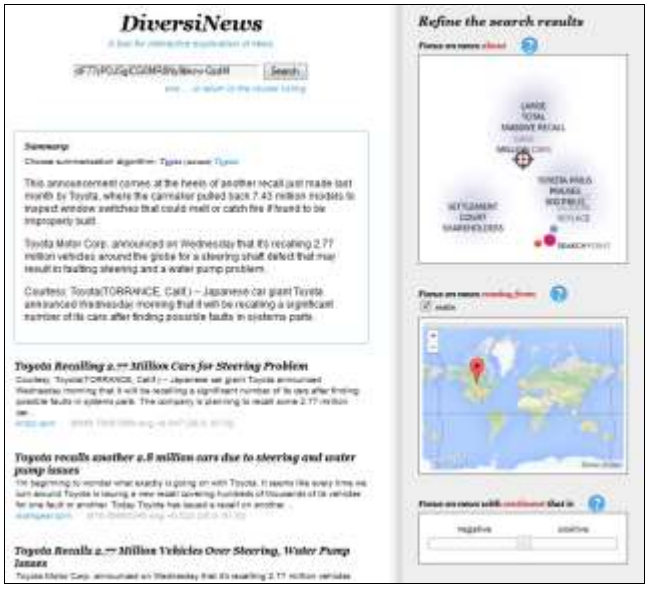

Figure 2. DiversiNews main interface.

3.2 iDiversiNews (iOS)

The iOS version allows near real-time browsing and analysis of news. Figure 1 gives an overview of the system architecture. The backbone of the server part are components implemented as web applications for the Apache Tomcat Java Servlet Container. Components are set up in a pipeline using Apache ActiveMQ, an open source messaging server.

Some functionalities, implemented in C++, have been exposed as (C++) web-services:

- SearchPont service – keyword extraction (k-means clustering) and projection of keywords and their relevance to a 2D plane.
- Google Summarizer integration.

In order to enhance user experience *iDiversiNews* server is to pre-processes as much data as possible before serving it to the client application. To achieve that purpose, data is downloaded (Figure 1, orange arrows) from *NewsFeed* and the *JSI Clustering Module* as soon as it is made available by those systems.

News Items are pushed through the Enrycher pipeline and stored in a repository.

News Clusters data is a list of cluster identifiers with corresponding news item identifiers. From this list an RSS-like structure is created by grouping news items from the repository. Those lists (clusters) are sent to the Descriptor service where all potentially useful metadata is extracted from single news articles, aggregated, calculated and added to the cluster itself. Metadata consists of the following: region, country, top tags, top categories, sentiment information, top entities mentioned, SearchPoint top keywords and their positions on the 2D plane, top image. News articles for each cluster are then ranked (Ranking Service) and a summary (Summarizer Service) is created according to that order. Each cluster is then indexed and saved to the repository.

At the end of the pipeline, a list of the biggest stories in the news is published through the Web API module and an aggregation of basic statistics is created. The client, in this case the iOS application, can request the latest list or make its own customized search. After selecting a cluster, all available metadata ("news story dimensions") is displayed and the user can then interact with the news story. User selection is sent back to the server (dashed arrows) where the cluster is re-ranked and a summary re-calculated.

Figure 3 shows the News Story Exploration Area. On the top an overview of the story is shown. Diversity controls

can be found on central part of the screen (geography, sentiment, topic); the summary and the ranked articles appear on the bottom.

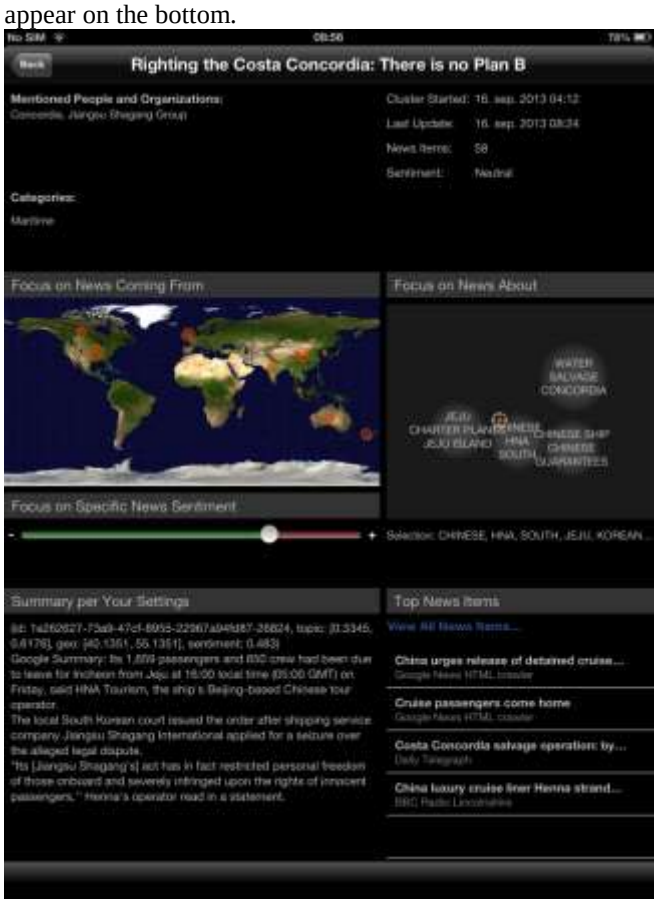

Figure 3: iDiversiNews

4 EVALUATION

Together with the RENDER project partners we performed a two-step evaluation of the DiversiNews web application. As a first step, we wanted to quantify fluency, informativeness and the impact of controls (the choice of topic and sentiment) on the generated summary. Next, we performed a user study with two domain experts from the Slovenian News Agency and other 14 non-expert users.

Impact of controls on summary. The evaluation was performed with two expert annotators on a random selection of 20 news clusters. Approximately 30% of the summaries were found to be fluent and informative. Regarding sentiment, the annotators were asked to mark 2 out of 8 summaries which have the most positive and most negative connotation, respectively. Sometimes polarity is not easy to detect, showed by lower recall (approximately 60%) compared to precision (approximately 75%). Topic-relatedness proved to be especially difficult to evaluate because of limitations in the user interface design. The annotators were asked to mark 2 out of 8 summaries which are most central with respect to the selected topic. The summarizer achieved approximately 90% F1 score on topic relatedness.

User Study. The user study was conducted according to three dimensions: a static evaluation, an interactive evaluation and a perceived utility evaluation. The **static evaluation** aims at assessing how self-explanatory the DiversiNews interface is. The results show that the majority of subjects found the interface very clear and self-explanatory from the very first moments of usage, and correctly identified the function and the behaviour of all the components. In the **interactive evaluation** the users were actually working with the system. ~81% of the subjects was either very pleased or pleased with the response time of the interface. The **perceived utility evaluation** aims to understand the real potential of DiversiNews as a platform for diversity aware news browsing. The subjects found summaries to be effective in capturing and representing relevant information. Moreover, the application succeeds in modelling different dimensions of diversity.

5 CONCLUSION

In this paper we described two applications – web and iOS – part of the RENDER news analysis case study. The applications allow users to explore news articles along several uncommon dimensions, diversifying them and discovering new aspects of a news story. The iOS application works in real time, making for a novel alternative to classic news reader apps. The web application was evaluated bot quantitatively, form the point of view of the impact of controls on the generated summary, as well as within a user study with domain experts and general users, showing a need for such diversity-aware solutions.

Acknowledgements

The research leading to these results has received funding from the Slovenian Research Agency and the European Union's Seventh Framework Programme (FP7/2007-2013) RENDER project under grant agreement no.257790.

References

[1] Rob Ennals, Beth Trushkowsky, John Mark Agosta, Tye Rattenbury, and Tad Hirsch. Highlighting Disputed Claims on the Web. *ACM International World Wide Web Conference (WWW)*, 2010

[2] Rob Ennals, Dan Byler, John Mark Agosta, and Barbara Rosario, What is Disputed on the Web? *4th Workshop on Information Credibility on the Web (WICOW)*, 2010

[3] Steinberger Ralf, Bruno Pouliquen and Erik van der Goot (2009). An introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): Information Access in a Multilingual World - *Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8. Boston, USA. 23 July 2009.*

[4] Linge Jens, Marco Verile, Hristo Tanev, Vanni Zavarella, Flavio Fuart, Erik van der Goot (2012). Media monitoring of public health threats with MedISys. *Living in Surveillance Socities.*

[5] Mitja Trampuš and Blaž Novak. Internals of an aggregated web news feed. *Proc. of SIKDD*, 2012.

[6] Tadej Štajner, Delia Rusu, Lorand Dali and Blaž Fortuna. Enrycher: service oriented text enrichment. *Proc. of SiKDD*, 2009.

[7] Jean-Yves Delort and Enrique Alfonseca. DualSum: a Topic-Model based approach for update summarization. *Proceedings of ACL*, 2012.