



RENDER
FP7-ICT-2009-5
Contract no.: 257790
www.render-project.eu

RENDER

Deliverable D6.4.3

RENDER Showcase

Editor:	Felix Leif Keppmann, KIT
Author(s):	Felix Leif Keppmann, KIT; Inga Shamkhalov, KIT; Flavio Fuat, JSI
Deliverable nature:	Report (R)
Dissemination level: (Confidentiality)	Public (PU)
Contractual delivery date:	30 September 2013
Actual delivery date:	30 September 2013
Suggested readers:	General public, people interested in DiversiNews or iDiversiNews,
Version:	1.0
Keywords:	Showcase, news aggregator, diversity news

Disclaimer

This document contains material, which is the copyright of certain RENDER consortium parties, and may not be reproduced or copied without permission.

In case of Public (PU):

All RENDER consortium parties have agreed to full publication of this document.

In case of Restricted to Programme (PP):

All RENDER consortium parties have agreed to make this document available on request to other framework programme participants.

In case of Restricted to Group (RE):

The information contained in this document is the proprietary confidential information of the RENDER consortium and may not be disclosed except in accordance with the consortium agreement. However, all RENDER consortium parties have agreed to make this document available to <group> / <purpose>.

In case of Consortium confidential (CO):

The information contained in this document is the proprietary confidential information of the RENDER consortium and may not be disclosed except in accordance with the consortium agreement.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the RENDER consortium as a whole, nor a certain party of the RENDER consortium warrant that the information contained in this document is capable of use, or that use of the information is free from risk, and accept no liability for loss or damage suffered by any person using this information.

Full project title:	RENDER – Reflecting Knowledge Diversity
Short project title:	RENDER
Number and title of work package:	WP6 – Dissemination, Exploitation, and Community Building
Document title:	D6.4.3 – RENDER Showcase
Editor (Name, affiliation)	Felix Leif Keppmann, KIT
Work package Leader (Name, affiliation)	Andreas Thalhammer, UIBK

Copyright notice

© 2010-2013 Participants in project RENDER

Executive Summary

Digital publishing of news items is common practise in today's interconnected world. Professional companies as well as individuals use the new media for the distribution of their content. Several different communication channels appeared, from news portals over blogs to micro blogging services. Users now are able to access and are confronted with a vast amount of content from all over the world. While this bears many new opportunities, the user faces also issues, which hamper these, for example caused by the effort for individuals to get a comprehensive overview and form an independent opinion about a news event.

Some opportunities are taken as an example:

- Independent of time and location: the reader is not limited to a specific location and the news media published in the region. Further, the reader may consume news at any time without being limited to specific opening hours or deliveries.
- Quick access: information about news events is distributed faster via new media than, for example, via a traditional newspaper. As soon as an author publishes a new item, the reader is immediately able to read the news item.
- Access to several opinions: from the independency of the location follows also that the reader is independent from the relatively small amount of local publishers. Thereby, the access to several more options about a news event is enabled, which allows the reader to form an own opinion on a broader base. Depending on the political system it may be the only way to read or publish independent news.

Some issues are taken as an example:

- Many communication channels: with the digitalization in the news area, several, partly new, communication channels have been established. Which channel is used, by which author blurs up and the reader has to adapt to too many different publishing formats.
- Amount of news items: the vast amount of content created by different authors. The time a reader has to spend on reading news item exceeds the benefit the reader is able or willing to spend.
- Incompleteness: considering the comprehensiveness that could be achieved in coverage of a news event, a selection of news items may be incomplete in terms of facts, opinions and viewpoints. Still, the number of available news items is higher, but it is more difficult to select the right sources to cover a news event as comprehensive as possible.

This is in short the setting for the RENDER showcase, a demonstration of technology developed within the project that combined in one application supports a user to cope with challenges of consuming digitally published news items. By reducing or solving the issues, we enable the user to read news items in a new way. An intuitive mobile user interface aggregates, ranks and summarizes the stream of new news items, according to the individual viewpoint of the user. This viewpoint is configurable by different aspects of diversity, e.g. keyword groups, geographical location, and sentiment. The content presented to the user will adapt as soon as the viewpoint is changed.

The following use cases have been derived and implemented in the showcase applications:

- Select News Story
 - Extended: Search News
- Read Summary
- Browse Ranked News

- Define Viewpoint
 - Derived: Select Keyword Group
 - Derived: Select Geolocation
 - Derived: Select Sentiment
- Explore Related Entities

Several components of the backend are hidden by the user interface. These have been developed or extended within the project and in combination take up the challenge to cope with the never-ending stream of news items appearing every second. Besides the crawling and aggregation of news items for the user, unique features distinguish these showcase application from already available systems, for example the extraction of several aspects of diversity, their combination and dynamic user-driven integration in the aggregation and ranking of content.

The components of the system can be classified by a layered architecture, including two separated processing pipelines.

- *Data Layer – Repository*: components responsible for the persistent storage of extracted and calculated data and metadata;
- *Processing Layer – Crawling, Extraction and Analysis*: components responsible for the crawling of news items, the extraction of content and analysis of different aspects of diversity;
- *Processing Layer – Ranking, Summarization and Access*: components responsible for the ranking and summarization of content, according to the user's request and for the accessibility of the system's functionality;
- *Presentation Layer – User Interface*: components responsible for an intuitive user interface.

Table of Contents

Executive Summary	3
Table of Contents	5
List of Figures	6
Abbreviations	7
1 Introduction	8
1.1 Scenario	8
1.2 Stakeholders	9
1.2.1 Government	10
1.2.2 Media	10
1.2.3 The Public	10
1.2.4 Enterprises	10
1.2.5 Individuals	11
1.3 Approach	11
1.3.1 Central News Access	11
1.3.2 Content Aggregation and Ranking	11
1.3.3 Diversity Revelation	11
1.3.4 Result Visualization	12
1.4 Use Cases	12
1.4.1 Select News Story	13
1.4.2 Read Summary	13
1.4.3 Browse Ranked News	13
1.4.4 Define Viewpoint	13
1.4.5 Explore Related Entities	14
2 DiversiNews	15
2.1 Development	15
2.2 Architecture	16
2.2.1 Overview	17
2.2.2 Layers and Component Groups	17
2.3 Components	18
2.3.1 Data Layer – Repository	18
2.3.2 Processing Layer – Crawling, Extraction and Analysis	18
2.3.3 Processing Layer – Ranking, Summarization and Access	22
2.3.4 Presentation Layer – User Interface	23
3 Demo Materials	34
3.1 Showcase Application	34
3.1.1 System Requirements	34
3.1.2 Setup	34
3.2 Dissemination Materials	34
3.2.1 Website	34
3.2.2 Screencast	34
3.2.3 Flyer	35
3.2.4 Poster	35
4 Conclusions	36
References	37
Annex A DiversiNews Web App	38
Annex B SiKDD 2013 Submission	42

List of Figures

Figure 1: Use Cases	12
Figure 2: DiversiNews Web Interface	15
Figure 3: Early iDiversiNews Mockup	16
Figure 4: Showcase Architecture	17
Figure 5: NewsFeed Demo Visualization	19
Figure 6: Enrycher Demo Visualization.....	20
Figure 7: SearchPoint Demo Visualization.....	21
Figure 8: iDiversiNews – News Overview	24
Figure 9: iDiversiNews –News Stories	26
Figure 10: iDiversiNews – News Story Details	28
Figure 11: iDiversiNews – Ranked News Items	30
Figure 12: iDiversiNews – News Item	32
Figure 13: DiversiNews – News Stories (cp. Figure 9)	39
Figure 14: DiversiNews – News Story Details/Ranked News Items (cp. Figure 10 and. Figure 11).....	40
Figure 15: DiversiNews – News item (cp. Figure 12)	41

Abbreviations

API	Application Programming Interface
DMOZ	from Directory.MOZilla.org, now Open Directory Project (ODP)
NLP	Natural Language Processing
POS	Part of Speech
RSS	Rich Site Summary

1 Introduction

Digital publishing of news items is common practise in today's interconnected world. Professional companies as well as individuals use the new media for the distribution of their content. Several different communication channels appeared, from news portals over blogs to micro blogging services. Users now are able to access and are confronted with a vast amount of content from all over the world. While this bears many new opportunities, the user faces also issues, which hamper these, for example, caused by the effort for individuals to get a comprehensive overview and form an independent opinion about a news event.

This is in short the setting for the RENDER showcase, a demonstration of technology developed within the project that combined in one application supports a user to cope with challenges of consuming digital published news items. By reducing or solving the issues we enable the user to read news items in a new way. An intuitive mobile user interface aggregates, ranks and summarizes the stream of new news items, according to the individual viewpoint of the user. This viewpoint is configurable by different aspects of diversity, e.g. keyword groups, geographical location, and sentiment. The content presented to the user will adapt as soon as the viewpoint is changed.

Several components of the backend are hidden by the user interface. These have been developed or extended within the project and in combination take up the challenge to cope with the never ending stream of news items appearing every second. Beside the crawling and aggregation of news items for the user, unique features distinguish these showcase application from already available systems, for example, the extraction of several aspects of diversity, their combination and dynamic user-driven integration in the aggregation and ranking of content.

The later of this section will motivate the need for new ways to cope with the vast amount of news items in the digital world, differentiate stakeholders interested in this issue, present the approach chosen in the project and explains the use cases, which have been the starting point for development. Section 2 presents DiversiNews, the application that integrates RENDER technology in one single product. A short explanation of the development steps is followed by an architectural overview and description of the integrated components. The mobile frontend iDiversiNews, presented at the end of the section, tackles all the use cases derived from the scenario before. At the end, section 3 gives a short overview about demonstration of the showcase, in particular the showcase website, dissemination materials and the requirements and setup of the showcase application. Section 4 concludes the showcase.

1.1 Scenario

Nowadays information technologies are more and more integrated in several areas of life and influence the habits of humans that take advantage of the new opportunities. The younger generations grow up in an interconnected world with computers, smartphones and social networks. Nearly every device is capable to retrieve and display information from all over the world. At the same time the amount of accessible information on the internet grows every day.

One example for this trend is the publishing of news and articles related to news events. In the past a shift to digital publishing has been observed. Publishers established online news portals aligned to their print media, solely digital news pages emerged and consumers incorporated the new media in their daily life. Not only professional publishers but also individuals are now able to write articles and publish these worldwide with only little effort. The border between journalists and amateur authors is blurring up, information about news events is often published initially by involved persons from their mobile devices or as a follow up at home on their computers. Furthermore both professionals and individuals may use different ways of publishing, for example news websites, blogs, micro blogs, or social networks. Each communication channel has its own requirements, technical limitations and user groups. Different ways of

publishing can be also combined or interlinked to broaden the readership, for example a link to an article on a news portal included in a micro blogging post.

A vast amount of content gets accessible via new media for an individual reader. They face new problems, e.g. information overload, but also new opportunities, e.g. access to different opinions. In the following some opportunities are taken as an example:

- Independent of time and location: the reader is not limited to a specific location and the news media published in the region. Further the reader may consume news at any time without being limited to specific opening hours or deliveries.
- Quick access: information about news events is distributed faster via new media than for example via a traditional newspaper. As soon as an author publishes a new item, the reader is immediately able to read the news item.
- Access to several opinions: from the independency of the location follows also that the reader is independent from the relatively small amount of local publishers. Thereby the access so several more options about a news event is enabled which allows the reader to form an own opinion on a broader base. Depending on the political system it may be the only way to read or publish independent news.

The digitalization of media and new ways of publishing news items also lead to new problems. Some issues are taken as an example:

- Many communication channels: with the digitalization in the news area, several partly new communication channels have been established. Which channel is used by which author blurs up and the reader has to adapt to too many different publishing formats.
- Amount of news items: the vast amount of content created by different authors. The time a reader has to spend on reading news item exceeds the benefit the reader is able or will to spend.
- Incompleteness: considering the comprehensiveness that could be achieved in coverage of a news event, a selection of news items may be incomplete in terms of facts, opinions and viewpoints. Still, the number of available news items is higher, but it is more difficult select the right sources to cover a news event as comprehensive as possible.

An interested and informed user may not be able to cope with these issues to take advantage of the new opportunities.

While it is not desirable to reduce and centralize the creation of news and thus hamper the opportunities emerged from increased diversity and richness of the new media, the question appears how new technology could support the reader in consuming news. A technological solution must tackle the issues mentioned while preserving or even enhance the opportunities.

1.2 Stakeholders

The most obvious beneficiary who would profit from an improvement of the situation described in the scenario above is the individual end user. On the one hand, every user of digital published news who is interested in a comprehensive, fast and complete overview of news will come across the issues of selecting, aggregation and reviewing different sources. On the other hand, the time to achieve this overview in the daily life of the user is limited.

Beside the individual user other stakeholders benefit of an improved situation in this scenario. A stakeholder thereby must not be a distinct person or organization but is in general a part of the society that is affected by a change of the situation. In the following a selection of different stakeholders and their

interest in an improvement of the current situation is described, in particular the external stakeholders: government, media and the public, as well as the end user stakeholders: enterprises and individuals.

1.2.1 Government

Democratic governments, in this context in particular the European Union, are interested in an independent and free access to news. Freedom of expression and media is a basic human right as well as a challenging task in the digital age. The EU High Level Group on Media Freedom and Pluralism claims a number of tasks and focuses of the European Union according the freedom of expression and media [5]:

- Ensure the un-discriminatory access to impartial news for the population for reflecting a wide range of opinions in news without an inadequate effort.
- Encouragement of pluralism in media in general.
- Warrant the dissemination of European news.
- Consider the changing media landscape in respect of development of new technologies and new business models.
- Ascertain that providers insure the transparency of the news service.
- Verify the neutrality of the news service provider and the mechanisms through which the news are transmitted.

The un-discriminatory access and the neutrality of mechanisms and content are obtained by providing a central and neutral search engine for all news available on the internet. The pluralism in media is ensured by coverage of the entire news amount on the internet and the capability to select between diverse opinions and sentiments. In order to give complete transparency as how individualized the news results depending on selected options are, the user is able to turn off the personalization.

1.2.2 Media

Diminishing or solving the issues of individual users in the digital world of publishing would have effect on the media itself. On the one hand it may help the individual author in investigations and research. On the other hand the transparency and comparability puts pressure on the media to achieve a higher quality of the content.

Furthermore the development of an adequate check and balances environment is promoted to prevent misuse of power based on capability to uncover partial news on the internet [4].

1.2.3 The Public

The general public is interested in having access to impartial news. Impartial news are the basis for free opinion formation of the general public. On the internet available news are public goods. Public goods are goods that are consumed for free. Hence, it is necessary to establish performance measurement criteria, because of absence of a monetary performance appraisal.

Encouragement of pluralism in media helps to establish a neutral and independent opinion formation of the general public and to avoid negative impact on democracy through falsified reporting.

1.2.4 Enterprises

Enterprises as stakeholders profit from a better access to news items as the consequence of better exploitation of available news on the internet and precise assessment of economic facts. News among others are raw materials for economic analysis, forecasts and benchmarking resulting from more neutral and faster justification.

1.2.5 Individuals

Private person as a user has interests in quality evaluation of news, similar to the interest of the public. In addition, no private user can keep up with the overwhelming volume of news produced each day. Therefore, search engines are necessary to find and structuring the required news articles to enable an optimal exploitation of available information. Structuring the news can be fulfilled by selecting and refining these by opinions and sentiments. The summarization functionality facilitates the feasibility of handling the amount of news articles.

1.3 Approach

The research project RENDER focuses on the development of algorithms, infrastructure and applications to collect and manage various information sources, identify and extract aspects of diversity and to present and visualize diverse information in a human-friendly manner. Several developed technologies have been combined in one application which improves the situation of a user in the scenario described above. It has been developed and deployed as part of a use case and integrates a meaningful subset of the RENDER architecture. Components of all architecture layers are incorporated and are accessible by a user friendly interface.

From a user perspective, the application supports the user in the process of accessing news articles, revealing different aspects of diversity and aggregation of the content. It solves or weakens the issues described in the scenario above. The main idea behind the application is not to relieve the user from forming a view on a particular news story or assess the value of articles, but to give easy access to different aspects of diversity and different viewpoints and at the same time provide references to the original articles.

In the following four important areas of functionality are described that have been identified to support the user in an appropriate way. While these describe general requirements for the application, the next section 1.4 describes specific use cases that guided the development.

1.3.1 Central News Access

Central news access includes the pre-processing of raw data published in diverse formats on the internet (e.g. diverse news from various news departments, blogs, wikis etc.). The crawled news are discovered, indexed and stored for later analysis. Information about and content of the original source should be preserved for correct referencing and detailed research. In the end the user has a central access point for news while still being able to read the original articles.

1.3.2 Content Aggregation and Ranking

Content aggregation and ranking tackles the issues emerging from the vast amount of content that are face the user in many cases. News items should be ranked, selected and aggregated by appropriate mechanisms according to the preferences of the user, for example with ranked news item lists or summaries. Not only news items but also content snippets may be selected that are of importance to the user. In all cases the original sources should be still accessible.

1.3.3 Diversity Revelation

Diversity revelation should support the user in forming an own opinion about a news event. This functionality is intended to refine, select and sort news items according to different aspects of diversity e.g. location, positive/negative sentiment or topic. In combination with adapted ranking and aggregation mechanisms it will help the user to reveal different aspects of diversity in the news. Therefore appropriate algorithms have to be developed to extract these metadata.

1.3.4 Result Visualization

Result visualization is important from the perspective of a user. The preceding functionalities may become useless if their results are not presented intuitive and easy accessible to the user. Both, the search and selection of news items as well as the configuration of different aspects of diversity should be integrated in a single frontend. The frontend enables users to access the news service, to perform searches, to use news overviews and to access search results.

1.4 Use Cases

A user in the situation described in the scenario in section 1.1 can derive several specific use cases for an application with respect to the chosen approach in section 1.3. All these use cases support the high level goal of a user to get a comprehensive, fast and complete overview of news. In the following the use cases that are supported by the developed application are described in detail.

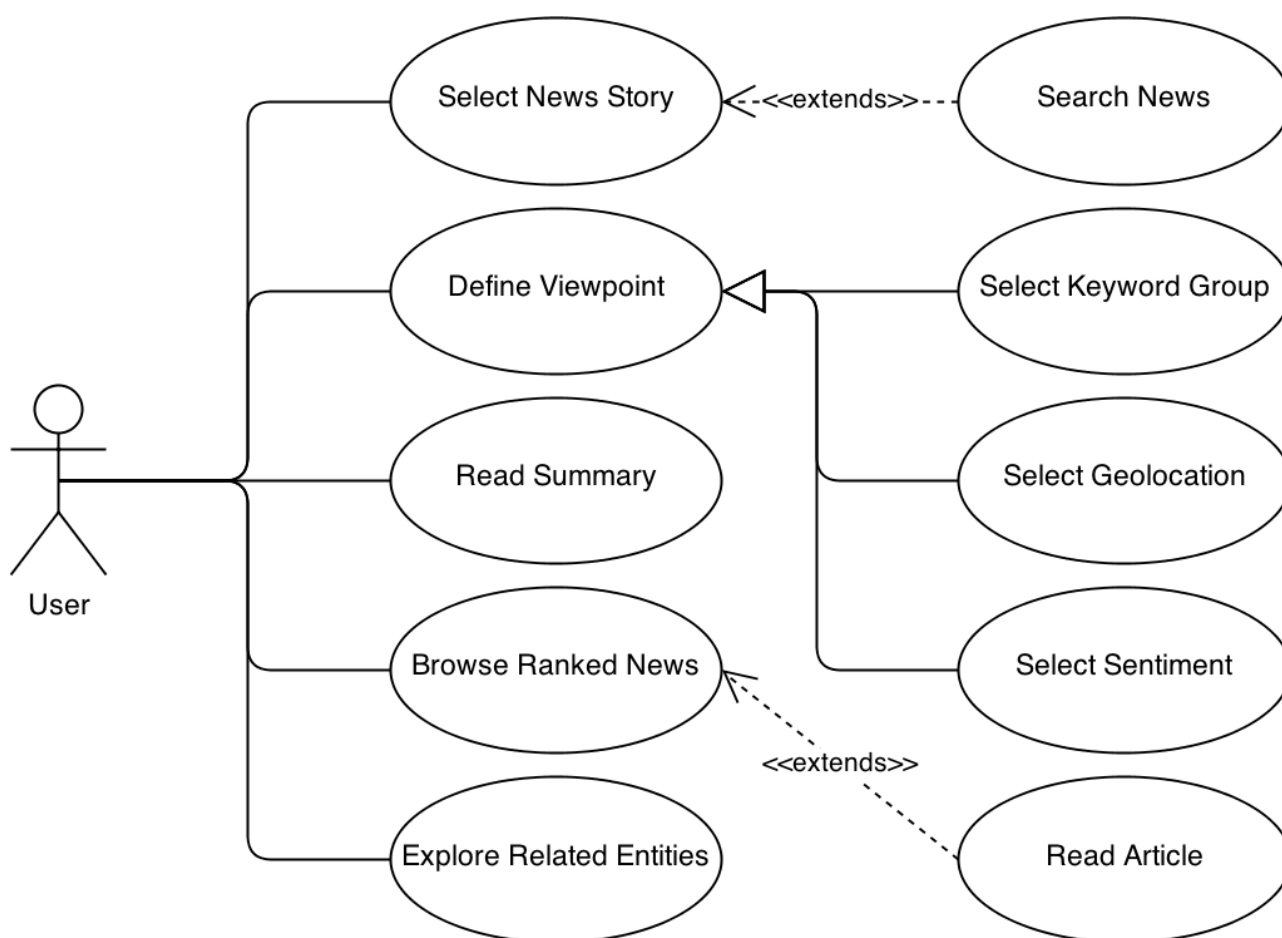


Figure 1: Use Cases

The use case diagram in Figure 1 visualizes in a comprehensive manner all these selected use cases. Implemented as an application this will cover the centralized access to news by selecting and searching for news stories, the aggregation and ranking of content by summarization and appropriate ranking of selected articles, the flexible definition of a viewpoint on the content and the presentation of results as well the access to the original sources. In the following each use case is described in more detail.

1.4.1 Select News Story

The user should be able to select a news story. A news story hereby is a collection of articles about the same news event, thus news stories are high level aggregated news. Currently available news stories should be listed and ranked by default by common measure, e.g. number of articles in a news story combined with the topicality of the articles. Different sorting and filtering mechanisms may help the user to select an interesting news story.

Search News Story

In addition to the listing and selection of news stories, the user should be able to search for particular news stories by providing for example keywords. As shown in Figure 1 this use case extends the use case for selecting a news story as the search functionality is an additional help for selecting the right story. The system must be able to select a meaningful subset of available news stories, rank these by relevance and display this pre-selection to the user. The user then may choose the right news story or adjust the search.

1.4.2 Read Summary

While the prior use case focuses on the selection of the right news story, this and the following use cases assume that a news story of interest has already been selected by the user.

To be able to get a fast overview, the system should provide a meaningful summary of the content of the articles in the news story. Summaries assist a user to manage the enormous amount of information and obtain an overview over the published news. The summary should figure out the most important facts concerning the topic in as few as possible sentences.

Furthermore, by reading the summary a user would like to get an overview over different polarities of the same event. Therefore the summary should be adjusted according to the defined viewpoint described in the use case in subsection 1.4.4.

1.4.3 Browse Ranked News

Beside the summary the system should provide a list of all articles grouped in the selected news story. The list should provide sufficient information per listed article to get a notion of the content on the first look, for example by showing the headline or a text snippet of each article.

Similar to the summarization described in the prior use case, the articles in the list should be ranked according to the defined viewpoint described in the use case in subsection 1.4.4. Thus more relevant articles are displayed more prominent, with respect to the defined viewpoint.

Read News Item

The purpose of the ranked list of articles is primary for the selection of the most relevant articles for the user. As a consequence a user should be able to also read the actual content of articles. Therefore the browse ranked articles use case is extended by a read article use case. The system should be able to provide this content whenever the user selects an article.

1.4.4 Define Viewpoint

While the prior use cases may be covered already by existing applications, the following use cases go beyond state of the art applications. These use cases, also in combination with the prior ones, will show achievements of the project and their potential implications on productive applications.

A major concern motivated in the scenario in section 1.1 is not only the access to the vast amount of information, but also the inability of the user to extract all the different opinions and facts about a news

event. Thus important opportunities emerged from the digital publishing may be hampered, for example by time constraints of the users.

The system should support the user with this task. It should enable the user to easy and fast extract different viewpoints on a particular topic. Different aspects of diversity should be made accessible and configurable by the application. Modified by the user, the summary and ranking described in the use cases above should adapt to the new configuration. For the showcase three aspects of diversity have been selected: topic focus, geographic focus and sentiment focus of articles. As shown in Figure 1 for each aspect a use case has derived. Combined this use cases serve the definition of the viewpoint of the user.

Select Keyword Group

The underlying news event of a news story may be very delimited, but can also be confusing complex and hard to understand. The more broaden the news story covers a news event, the less a user is able to extract and aggregate the information about the different parts of the story. The system should support the user and provide a way to focus on particular parts of the news story. Therefore the most important keyword groups should be extracted and presented to the user. The selection of a particular keyword groups is then reflected in the ranking and summarization of the news items.

Select Geographic Focus

In a globalized world with news agencies in every county, the authors and publishers write and publish content with their specific national and cultural background. To be able to read about a news event not only from sources with a familiar background, but also from the perspective of humans on other locations on the earth, is an important opportunity enabled by digital news publishing. The user should be able to weight the importance of sources by their geographical location and thereby be able to easy access these different perspectives.

Select Sentiment Focus

Depending on the actual topic of the news story, the authors of articles may discuss in a controversial manner with many positive and negative arguments, or unanimously argue in one direction. In both edge cases, but also in general, it is important for an interested and informed user to asses both, the negative and the positive arguments to form an independent view on the topic. Therefore the system should enable the users to weight the articles by sentiment on a negative to positive scale. In the case of controversial topics it may be used to distinguish and aggregate the plenty of arguments. In the case of uniformed argumentations in most of the articles it may help the user to find the few contradicting arguments.

1.4.5 Explore Related Entities

The focus on a single news story may narrow the view of a user to much and thus hamper the ability of the user to explore the full context of the event. This may counteract this effect by providing related entities. Related entities should be entities that have a direct relation to entities or parts of the news story, but are not in the focus of the news story. They may guide the user to related topics and thus enable to explore the context of the news even.

2 DiversiNews

In the following the software solution is presented that has been developed to tackle the issues motivated in introduction by implementing the use cases described in section 1.4.

2.1 Development

The development of DiversiNews has been documented in the deliverables D3.2.1 (Diversity-aware summarization), D5.2.2 (Diversified News Service), D5.2.3 (Refined Diversified News Service), D5.2.4 (Evaluation of the diversified news service) and this document.

DiversiNews
A tool for interactive exploration of news.

crisis Search

««« ... or return to the cluster listing

Summary:
Choose summarization algorithm: *Type1* (current) *Type2*

"Although this is a smaller rise than recorded in recent months, and there are good reasons to expect long-term unemployment to stop rising in coming months, there are now almost 900,000 people in the UK who have been out of work for more than a year.

That's why we've committed to supporting the hardest-to-help people over a two-year period through the Work Programme so that we can help them overcome their barriers to work and get them into sustainable jobs."

The figures were published on the same day that the Bank of England downgraded its growth forecast for 2013 to around 1% and warned that output would remain below pre-financial crisis levels for the next three years.

The World Failed Sri Lanka. And Continues to Do So.
The top story buzzing around the UN today the soon-to-be-released report on the failure to protect civilians caught in the final days of the Sri Lankan civil war. Parts of the report were already leaked to the BBC, and Ban Ki Moon is expected to make it p...
[undispatch.com](#) (105 69775812 eng -0.638 [0.0,0.0])

Fiscal Cliff? Obama Urged Not To Panic
WASHINGTON -- Republicans and deficit hawks are raising unnecessary alarm over the so-called "fiscal cliff" to pressure President Barack Obama into a "grand bargain" he shouldn't make, progressive economists and scholars said Tues...
[huffingtonpost.com](#) (123 69783042 eng -0.000 [37.2,-95.8])

Petraeus scandal creates 'nightmare' for Mount Laurel lawyer and family
"I've had no sleep, to be honest with you," Khawam, 40, said at his Camden County office on Tuesday. Before discussing his shift from the quiet life of a suburban New Jersey attorney to the object of inquiries from everyone "from CNN ...
[phillyburbs.com](#) (179 69807801 eng +0.075 [38.0,-97.0])

Refine the search results

Focus on news **about** ?

OBAMA
STRIKE
REPUBLICANS
UNION
CLIMATE
PROTESTS
PETRAEUS
KELLEY
WORD
INFLATION
INFLATION
PARTIES

Focus on news **coming from** ?

☐ enable

Google

Terms of Use

Focus on news with **sentiment that is** ?

negative positive

Figure 2: DiversiNews Web Interface

The showcase is based on one of Google, one of the members of the project. During the development was focus on the development and integration of backend components while user interface has been developed as simple web application. Nevertheless, the web frontend shown in Figure 2 already contains all major functionality and has been used for a first evaluation in deliverable D5.2.3 and for the final

evaluation in D5.2.4. Further details of the web application in comparison to the developed showcase application are shown in Annex A.

After the final evaluation of the diversified news service in delivered in D5.2.4, the user interface has been re-designed and implemented as a mobile application for Apple iPads. The showcase application can be seen as an additional development step towards a production-ready user interface. Figure 3 shows an early mockup of the interface design. Feedback from users and the results of the evaluation have been taken into account and influenced the final design presented in section 2.3.4. Still, the user interface may be developed further to incorporate user feedback.

The backend system of the showcase application, described in section 2.3.1, 2.3.2 and 2.3.3, has been partially refactored and improved, for example a re-written sentiment analysis module.

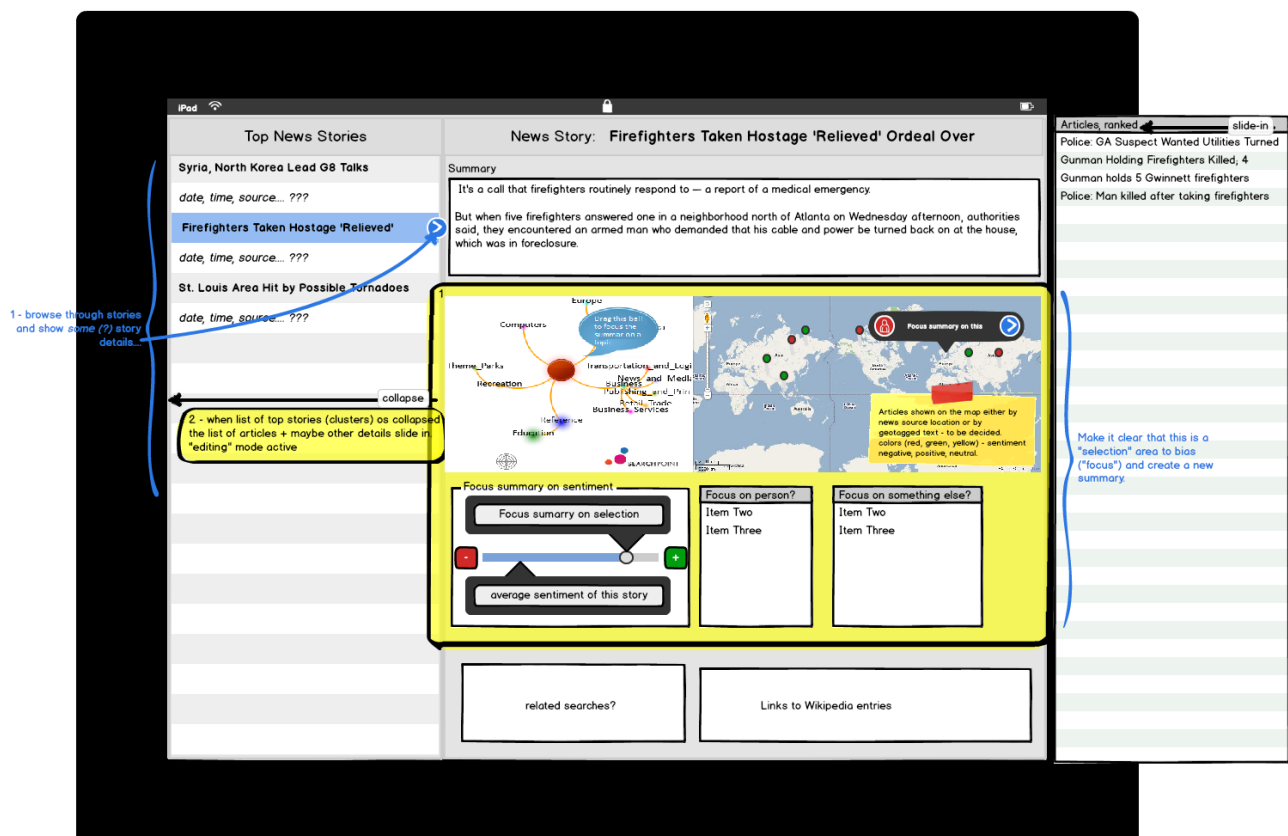


Figure 3: Early iDiversiNews Mockup

2.2 Architecture

The development described in the preceding section lead to a specific software architecture that fulfils the use cases derived from the scenario in section 1.4 and will be described in the following. Many of the components integrated in this architecture are part of the overall RENDER architecture, while some components have been developed specific for this application. In the following subsection 2.2.1 the overall picture of the showcase architecture is described while section 2.2.2 classifies the components by layers and function groups.

2.2.1 Overview

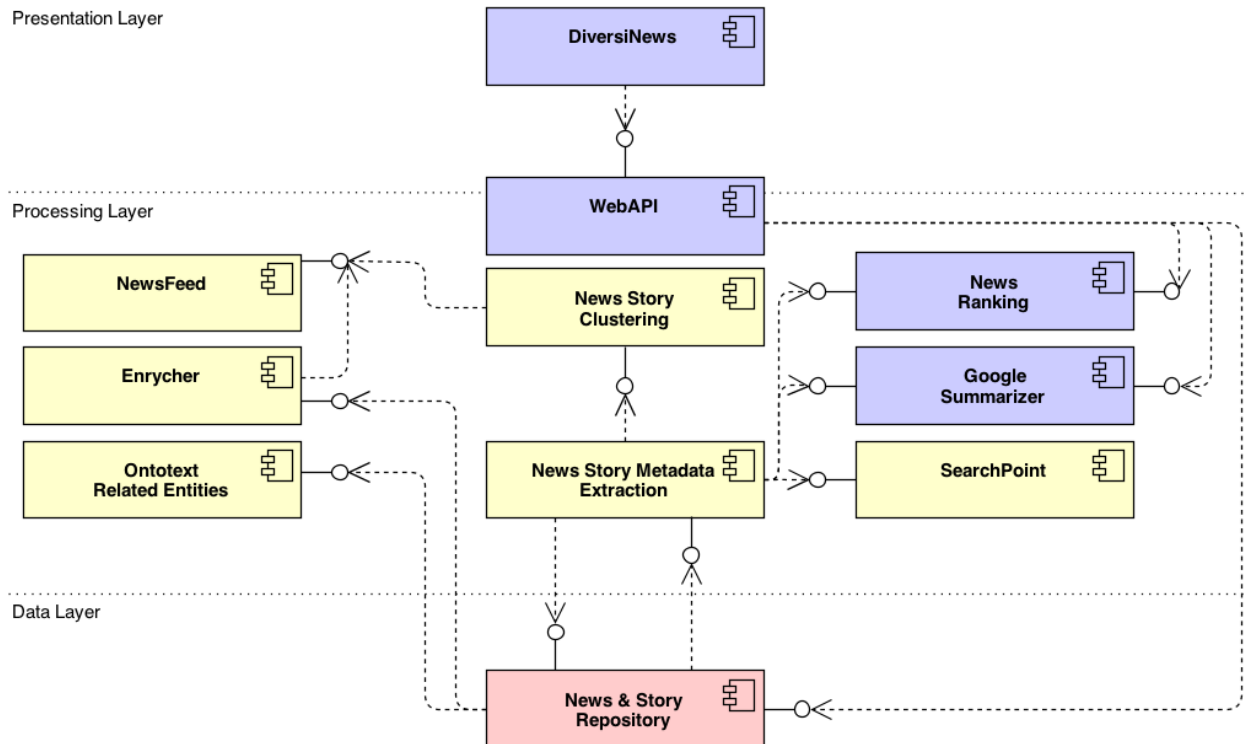


Figure 4: Showcase Architecture

A component diagram of the RENDER showcase architecture is shown in Figure 4. It consists of a classical layered architecture with a data layer for storage components at the bottom, a processing layer for all components covering extraction, processing and analysis of data in the middle as well as a presentation layer on top which handles all the interaction with end users or external applications. Each component is classified both by the position in one of the layers and by colour indicating the membership in a particular group of functionality. Both are explained in more detail in the following subsection.

The dashed arrows between the components indicate the data flow in form of API calls. A component may call the API of another component, indicated by the arrow, hand over data as arguments for the function call and may receive data as result of the function call. Thus the main data flow is contrary to the arrows in the diagram which shows the API calls. Components must not reside in the same system and/or location but are distributed and use corresponding communication techniques, for example web service or remote procedure calls. The components and their interactions are explained in detail in section 2.3.

2.2.2 Layers and Component Groups

Each component is classified by a layer and a function group within the showcase architecture. Three groups – visualized by colour – classify components by functionality. These groups depend partly on each other, but are usually executed independently. All components form a pipeline for the crawling, extraction, storage, ranking, summarization news and the presentation of results to as well as interaction with the user.

The yellow function group contains all components required to crawl, extract and analyse news article from all over the world. This part of the overall pipeline is executed as soon as a new news item is published on an external website and discovered by the system. All components of this group are classified as members of the processing layer. Components of this group are described in detail in section 2.3.2.

The red function group contains the repository component and thus the only component in the data layer. It contains all data extracted and processed by the extraction pipeline and supplies subsequent processing steps with enriched data. This component is the only permanent active part of the system. Components of this group are explained in detail in subsection 2.3.1.

The blue function group contains all components that are executed on demand per user. It contains all user interfaces, the access component as well as the components responsible for individual ranking and summarization. Components of this group are described in subsection 2.3.3 and 2.3.4.

2.3 Components

This section provides an overview of the components, their interaction and the data flow. The description level is on the level of functionality. More detailed technical description, deployment and system boundaries are provided by respective deliverables, among others D3.2.1 (Diversity-aware summarization), D5.2.2 (Diversified News Service), D5.2.3 (Refined Diversified News Service) and D5.2.4 (Evaluation of the diversified news service) may serve as a starting point.

2.3.1 Data Layer – Repository

The data layer consists of only one component, the article and story repository. The components are coloured in red in Figure 4 and are described in the following.

Article & Story Repository

The article and story repository is the central storage for all crawled and calculated data. On the one hand single enriched news items are stored and provided for further analysis. News items are processed by the crawling, extraction and analysis pipeline described in the next section. Both, the content of the news item as well as extracted metadata are stored.

On the other hand the repository stores and provides the information about news stories. News stories are groups of news items that have the same topic in common. When a new news item appears, for each news story that is changed or newly created, the components in the pipeline extract and calculate metadata that is stored along with the structure of the news story in the repository.

Beside the pure storage functionality the repository takes a more active role in the system and acts as a controller for the extraction pipeline described in the next section. For each crawled news item the repository coordinates the correct execution of the processing components.

2.3.2 Processing Layer – Crawling, Extraction and Analysis

The processing layer consists of two separate pipelines of components. The first pipeline, described in this section, is executed per newly crawled news item. It processes the articles, extracts and calculates metadata, aggregates news items to news stories and stores all results in the repository. The components are coloured in yellow in Figure 4 and are described in the following. The second pipeline, described in the next section, is executed per user request and handles the analysis according to the configuration requested by the user.

NewsFeed

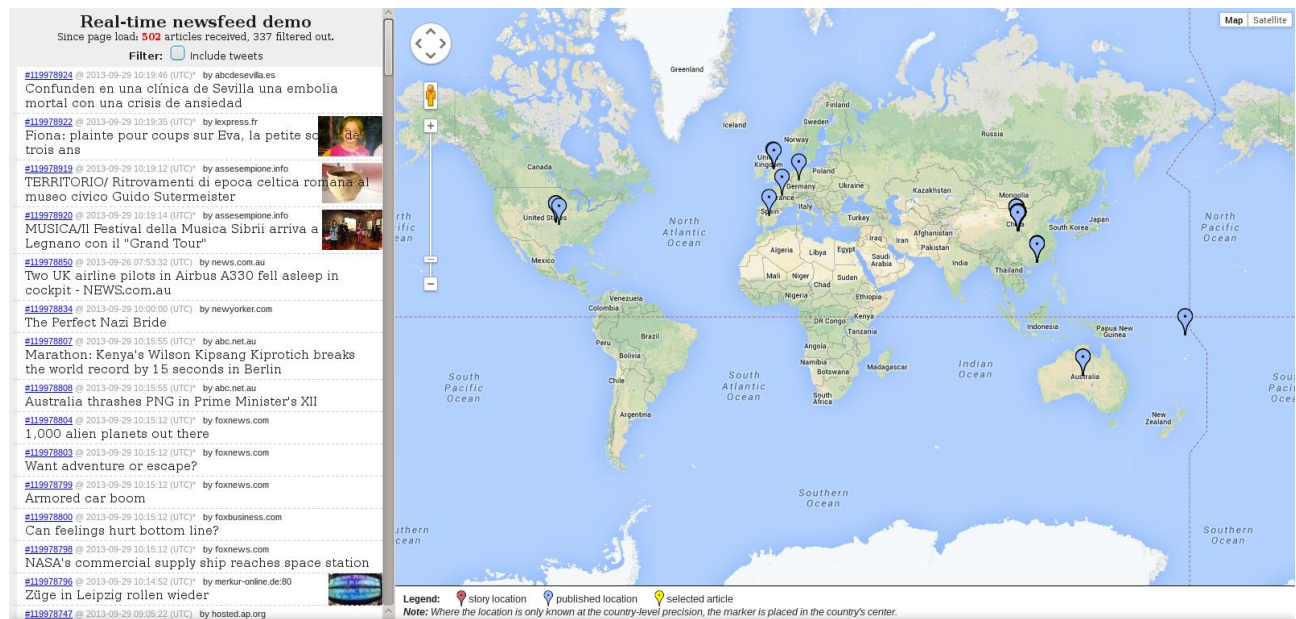


Figure 5: NewsFeed Demo Visualization

The NewsFeed¹ component is responsible for the supply of the system with new news items. Figure 5 shows a demo visualization² of the news feed. As soon as a news item appears, the map and list are updated and provide first information about the crawled news items. The component monitors several sources mainly RSS feeds of news websites from all over the world but also for example twitter feeds and blog posts. New news items are downloaded and first pre-processing steps are executed. These include content cleansing, association of a geographical location and language detection.

Content cleansing strips of non-content parts, e.g. markup languages, from the news item. The result should be the textual content of the news item. Association of a geographical location concerns the location of the publisher. Each publisher is associated with a geographical location which is either learned from public listings of news publishers that are crawled and include usually country and city, or as a fall-back mechanism are determined through heuristics on WHOIS queries. If available, the location is extracted from the WHOIS entry. Sources with national TLDs in the hostname are automatically assigned to the corresponding country. If no city is available, the publisher's location is mapped to the geographical centre of the country. A more detailed description of the NewsFeed component is provided by [2].

As soon as news item appears and is pre-processed, the system gets notified and subsequent components retrieve the content and already extracted metadata from the NewsFeed.

Enrycher

¹ <http://newsfeed.ijs.si/>

² http://newsfeed.ijs.si/visual_demo/

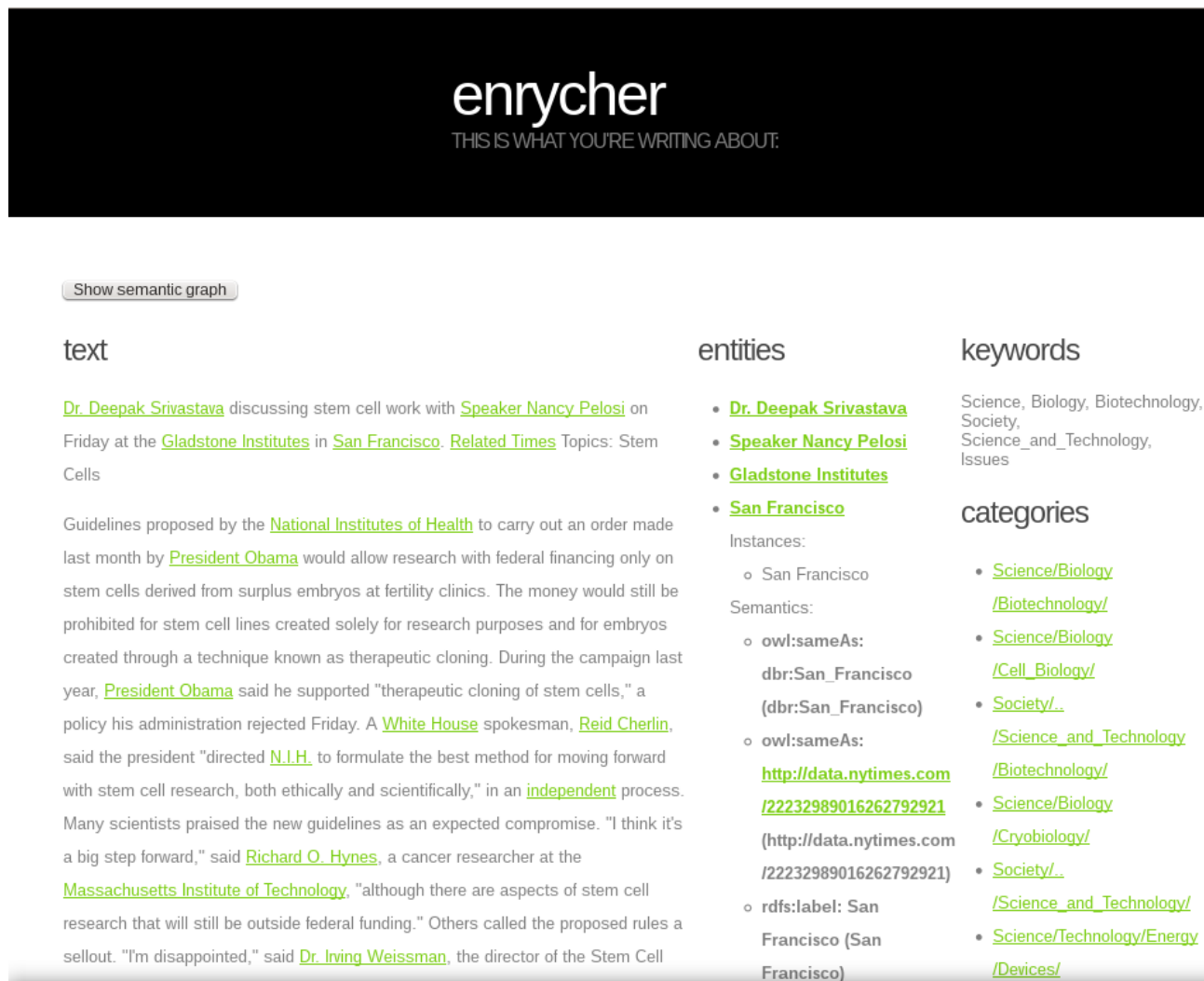


Figure 6: Enrycher Demo Visualization

Enrycher³ is the main natural language processing (NLP) component of the system. For each new news item the news and story repository requests an enriched version of the content. It performs part of speech (POS) tagging, sentiment analysis, named entity extraction and resolution and DMOZ classification. In particular the sentiment analysis module in Enrycher has been improved within the project. A supervised method is now being used with significantly improved performance. Enrycher is described in more detail in [3].

For every crawled news item the news and story repository requests an enriched version. The Enrycher component gets the clean text from the NewsFeed component and processes it in its internal NLP pipeline. The annotated text is afterwards stored in the repository.

Ontotext Related Entities

Every news item is stored in the repository is in a second step attached with a set of related entities. Therefore an external service developed and hosted by Ontotext is called. Related entities are retrieved and stored along with the existing metadata of the news item in the repository.

News Story Clustering

³ <http://enrycher.ijs.si/>

A stream clustering method has been developed 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. A new news item is inserted into the cluster with the nearest centroid; if the nearest centroid is farther than some threshold, the news item seeds a new cluster. Distances are defined via the cosine similarity measure. Individual news item' weight/contribution to the centroid is attenuated exponentially to prevent old stories from lingering in the system for too long. Overly old news items are discarded. Each time a news item is added to a cluster, the cluster's cohesion is evaluated; if it is too low, the cluster is split in two by recursive clustering. Periodically, a few iterations of k-means are run with existing centroids as the starting point to prevent the clustering quality from deteriorating too badly.

SearchPoint

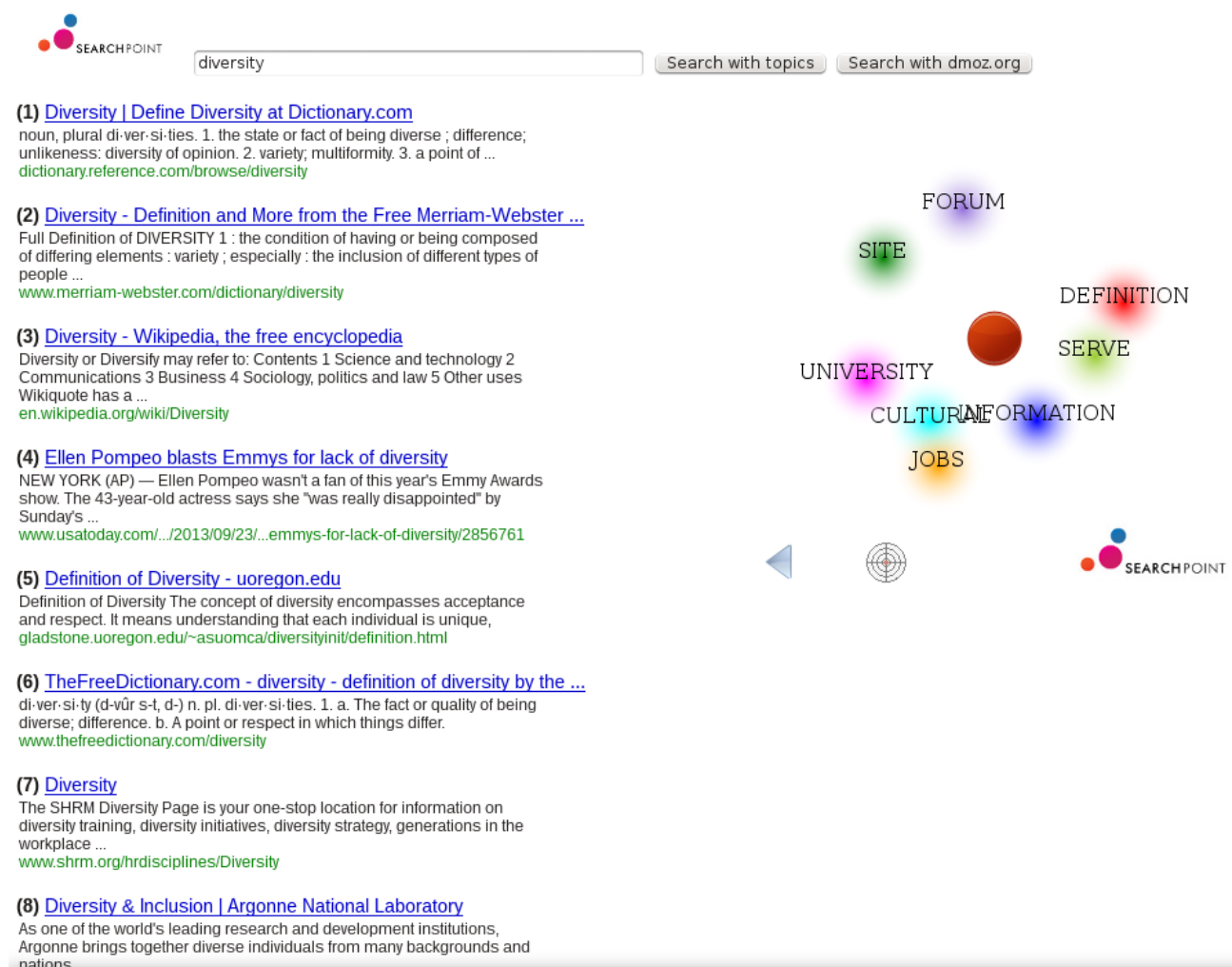


Figure 7: SearchPoint Demo Visualization

From every news story the main keywords are extracted and grouped in at most five groups (clusters). Groups of keywords are projected to a two-dimensional plane to be later displayed on the user interface. SearchPoint⁴ is used for the clustering as well as the projection of keywords on a 2D plane via a widget in the presentation layer. Figure 7 shows a demo visualization of SearchPoint results.

News Story Metadata Extraction

⁴ <http://searchpoint.ijs.si/>

The news story metadata extraction component ties all analysis steps for news stories together. First, for every new news item the repository requests a recalculation of then news stories. Therefore the clustering component is called and a recalculation of the news stories is requested.

Next, for all news stories that are created or changed, all potentially useful metadata from the news items of the respective news story is extracted, aggregated, calculated and added to the news story metadata in the repository. 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 and top image.

News items for each news story are then ranked and summarized according to that order. Each news story is then indexed and saved to the repository. The default ranking and summary are generated to speed up the first requests of users which contain usually default values for all configuration options.

2.3.3 Processing Layer – Ranking, Summarization and Access

In contrast to the first pipeline described in the preceding section, which is executed as soon as new news items are crawled, the second pipeline is executed per user request and handles the analysis according to the configuration requested by the user. The components are coloured in blue in Figure 4 and are described in the following.

WebAPI

The backend system, including all components of the data and processing layer, is accessible via a WebAPI component. Thereby external applications are able to access the functionality of the system via a defined interface as a web service. In the showcase application the mobile app described in section 2.3.4 acts as a lightweight frontend. All user requests are delegated to the WebAPI, processed in the backend system and the results are send back and presented in the app.

Internal the WebAPI accesses the data stored in the repository on the one hand to request a list of top news stories or to perform a keyword search on the indexed news stories. On the other hand, once a news story has been selected, it requests a new ranking and a new summary of the ranked articles according to the configuration of the user. Both ranking and summarization are described in the following.

News Ranking

On demand by a user request or in the news story metadata extraction, a ranking of the news items of a selected news story is performed by the news ranking component. A score is assigned to each news item. The score represents the distance of the news item from a user selection among three aspects of diversity: keyword group, geographical location and sentiment.

All three score vectors (keyword scores, geographical distance, sentiment) are normalised to have a mean zero and standard deviation one, thus making them comparable. For each news item the three scores are summed up (keywords score + geographical distance score + sentiment score). News items are sorted by the sum of the scores, allowing the display of more relevant items first.

For the keyword score each keyword has its own weight based on the significance in the cluster. Then, based on the user selection, a weighting factor is assigned to each group of keywords. The factor is inversely proportional to the distance of user selection coordinates to keywords coordinates. Thus, each keyword gets a weight that is “keyword weight * cluster weight”. Top five keywords are found using the weighting method above. Each news article gets a weighting factor based on the Okapi BM25 formula⁵.

For the geographical distance score each news article gets a weighting factor based on the geographical distance from the selected point.

⁵ http://en.wikipedia.org/wiki/Okapi_BM25

For the sentiment score each news article gets a weighting factor based on the difference between the selection and article sentiment.

Google Summarizer

On demand by a user request or in the news story metadata extraction, a summary of the news items of a selected news story is performed by the Google summarizer component. The summary is calculated with respect to the ranking of the news items calculated before.

Two summarization algorithms were implemented in the scope of the project; the first was developed by Google [1] 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. For the showcase system only the Google summarizer is used.

2.3.4 Presentation Layer – User Interface

iDiversiNews

In the preceding sections both the data layer and the processing layer have been explained. With the WebAPI as a central interface to the system, applications in the presentation layer can be developed independent from each other and without modifying the backend system.

In the following iDiversiNews is presented, the mobile frontend app for iOS based devices, in particular iPads. The app is currently the only component in the presentation layer and from a user's perspective it is the only visible part that embodies the complete DiversiNews system. Requests to the backend system are transparent and not visible to the user.

While the backend has been developed to collect, extract and calculate all the data required covering the use cases mentioned in section 1.4; iDiversiNews has been designed to provide a user interface that integrates the search and selection of a news story, the individual configuration of a viewpoint and the visualization of the results in an intuitive manner. A user should be able to use the app without prior knowledge.

In the following all views and options of the app are described and applicable the respective use case is mentioned.



Figure 8: iDiversiNews – News Overview

Figure 8 shows the entry view of iDiversiNews. The top half consists of a tag cloud containing the most often mentioned topics in the current set of news items. With increasing number of appearances the text size of the tag grows. The bottom half visualizes on a world map the publisher's locations. As bigger a blue circle appears as more news items are published in the same region.

Both, tag cloud and world map are touch sensitive and lead to a pre-configured news overviews described below. Touches on a tag in the cloud lead to a search of the specific tag and touches on the world map lead to a news story overview filtered by distance to the touched location. A touch on the button below leads to a default unfiltered news overview. The button in the left bottom corner allows updates the overview.

The news overview already covered partly the "select news story" use case described in section 1.4.1. A user may already define a topic or region of interest.

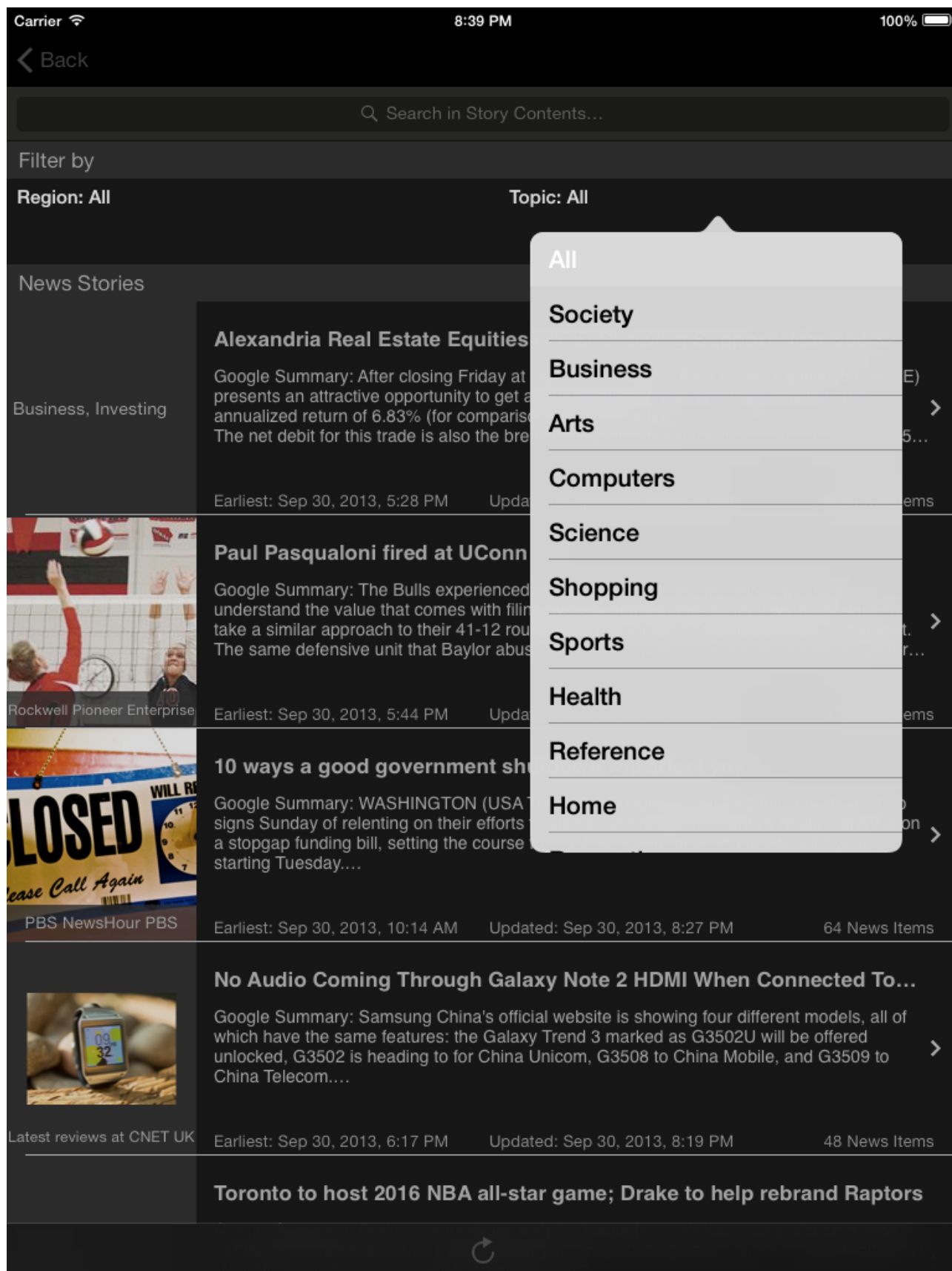


Figure 9: iDiversiNews –News Stories

Figure 9 shows the news stories overview. It shows a list of news stories related to current news events happening around the world. A user is able to perform a keyword search and will get a re-ranked and filtered list containing only relevant news stories, ranked by the “distance” to the keyword. Additionally a user may filter the list by region or topic. One filter may be already set if the user touched the tag cloud or world map at the entry view. Once a user touches a news story in the list, the news story details view is shown.

The news overview covers the “select news story” and “search news” use cases described in section 1.4.1.

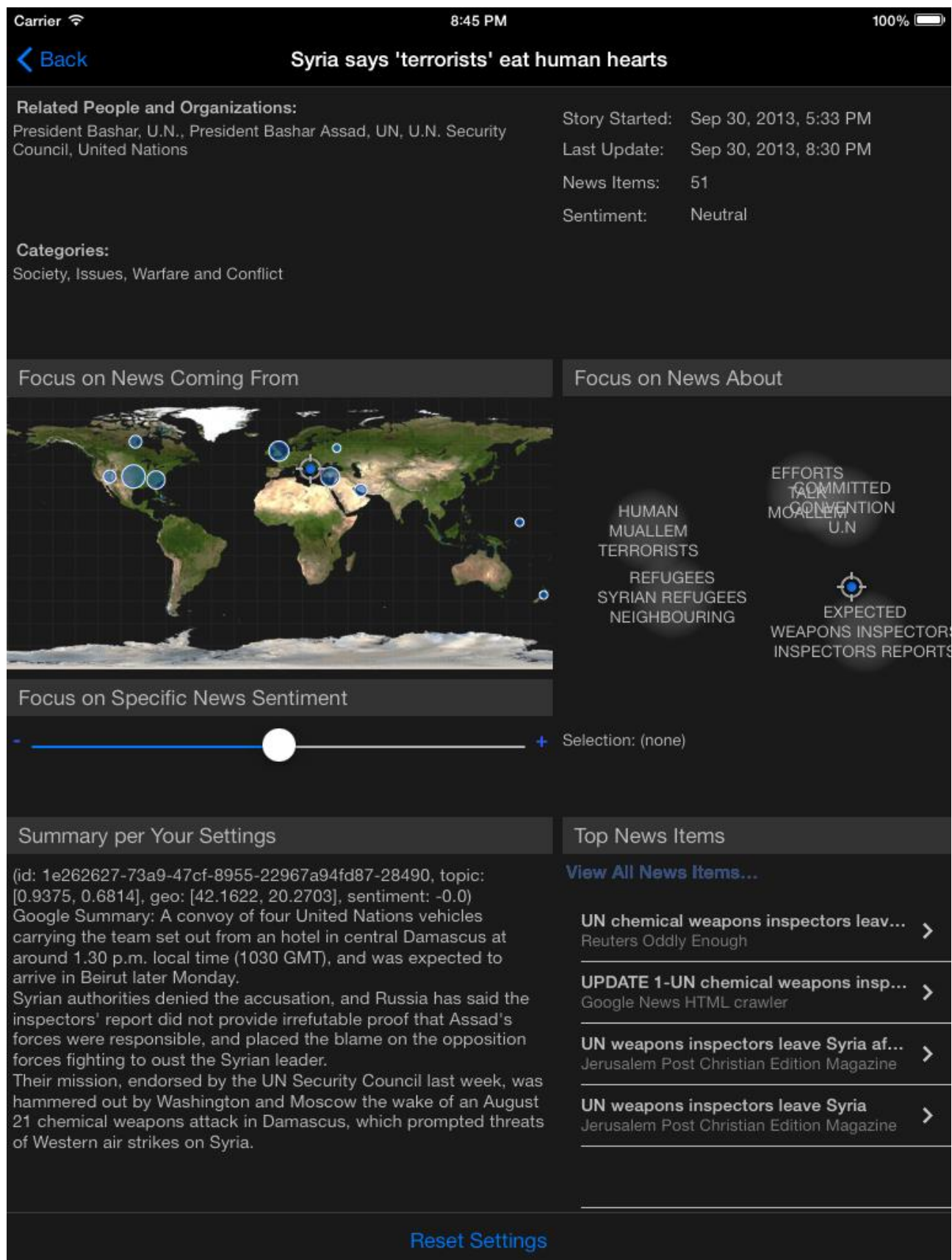


Figure 10: iDiversiNews – News Story Details

Figure 10 shows the detailed view of a news story. In the top part different metadata characterizes the news story. Related people and organization hint the user to potentially interesting related topics, the category classification is shown and the start and end time as well as number of news items and average sentiment are mentioned. It should help the user to pin down the news story.

In the middle part the configuration panels for the three aspects of diversity is displayed. The world map shows the geographical location of news items of the news story. As bigger a blue circle appears as more news items have been published in the particular region. A user may choose the focus from a geographical point of view by moving the cross to a specific region of the world. By moving the sentiment slider below the map a user may choose the focus from a sentiment point of view in the range from negative to positive. In the 2D pane on the right side visualizes the extracted keyword groups. A user may adjust his viewpoint by moving the cross one of the keyword groups. Summary and ranking of the news items below will be adapted as soon as the configuration of one of the options changes.

The middle part of the news story details view covers the “define viewpoint” use case with the derived use cases “select keyword group”, “selection geographical location” and “select sentiment” described in section 1.4.4.

In the bottom part of the view contains a summary as well as the top news items of the news story. Both are dynamically updated when one of the configuration parameters is changed. A re-ranked list and an adapted summary requested from the backend and displayed in the news story details view. If the user touches a news item, the respective news item will be shown in the news item view described below. A touch on the “view all news items...” button leads to the next view with the complete list of ranked news items.

The bottom part of the news story details view covers the “read summary” use case described in section 1.4.2 and partly already the “browse ranked news” use case described in section 1.4.3 because the top ranked news are displayed.

Carrier 8:45 PM 100%	
Back Syria says 'terrorists' eat human hearts	
Society, Issues, Warfare and Conflict, Specific Conflicts, Weapons Reuters Oddly Enough	Sep 30, 2013, 6:20 PM UN chemical weapons inspectors leave Syria BEIRUT (Reuters) - U.N. chemical weapons inspectors investigating allegations of chemical and biological weapons use during Syria's civil war left Damascus on Monday after their second mission in two months, witnesses said.
Society, Issues, Warfare and Conflict, Specific Conflicts, Weapons Google News HTML crawler	Sep 30, 2013, 8:28 PM UPDATE 1-UN chemical weapons inspectors leave Syria BEIRUT, Sept 30 (Reuters) - U.N. chemical weapons inspectors investigating allegations of chemical and biological weapons use during Syria's civil war left Damascus on Monday after their second mission in two months, witnesses said.
Society, Warfare and Conflict, Issues, Specific Conflicts, War on Terrorism, United Nations, Multilateral Jerusalem Post Christian Edition...	Sep 30, 2013, 6:20 PM UN weapons inspectors leave Syria after second mission United Nations chemical weapons inspectors in Syria to investigate allegations of chemical and biological weapons use during the country's ongoing civil war departed from Damascus on Monday, witnesses said, ending their second mission in two months.
Society, United Nations, Multilateral, Government, Chemicals, Business, Missions, Weapons Jerusalem Post Christian Edition...	Sep 30, 2013, 6:20 PM UN weapons inspectors leave Syria United Nations chemical weapons inspectors in Syria to investigate allegations of chemical and biological weapons use during the country's ongoing civil war departed from Damascus on Monday, witnesses said, ending their second mission in two months.
Society, Chemicals, Business, Weapons, Warfare and Conflict, Issues, Multilateral, Government Novinite	Sep 30, 2013, 8:28 PM EU Says Ready to Help in Elimination of Syria's Chemical Arsenal Michael Mann, a spokesman for European Union foreign policy chief Catherine Ashton, has told reporters in Brussels that the bloc is yet to receive an official request from the organization.
Society, Issues, Peace, Activism and Peace Work, Regional	Sep 30, 2013, 8:28 PM Bulgaria Welcomes UN Resolution on Syria's Chemical Weapons "Putting chemical weapons in Syria under international control, their subsequent destruction, ensuring transparency and guaranteeing their non-use in future is a significant step towards eliminating this threat to the Syrian people and to the countries

Figure 11: iDiversiNews – Ranked News Items

Figure 11 shows the ranked news items of a news story. Once a user has defined a viewpoint in the news story details, this view provides the full list of news items in the news story. The news items are ranked according to the three aspects of diversity. For each of the news items the title, date and time of publishing, a text snippet and if available an image are shown. The user may touch a news item to read the content on the news item view. On the bottom the list may be updated on demand with news item that appeared in the meantime and are classified as members of the selected news story.

The bottom part of the news story details view covers the “browse ranked news” use case described in section 1.4.3.

Carrier 8:47 PM 100%

[Back](#)

Syrian FM claims some rebels eat human hearts

UNITED NATIONS Syria's foreign minister has claimed at the U.N. that his regime is fighting a war against al-Qaida linked militants who eat human hearts and dismember people while they are still alive, then send their limbs to family members.

Published On: Sep 30, 2013, 8:30 PM

Related People and Organizations:

U.N., Walid, Syrian National Coalition, President Barack Obama, U.N., Lebanon based Al Mayadeen TV, Al Moallem, President Bashar Assad, U.N. Security Council, U.N., U.N.

Categories:


Society, Issues, Warfare and Conflict, Specific Conflicts, Regional

CBSNews.com / CBS Evening News / CBS This Morning / 48 Hours / 60 Minutes / Sunday Morning / Face the Nation


CBSNEWS Video US World Politics Entertainment Health MoneyWatch SciTech Crime Sports » More [Log in](#)

Syria Crisis


EXPLORE: Latest / President Assad Interview / Foreign Intervention / Chemical Weapons / Humanitarian Crisis



Syrian FM claims some rebels eat human hearts



OPCW inspectors plan to smash Syria chemical weapons-making ability by start of




Assad says he will abide by U.N. resolution

[Play VIDEO](#)

CBS/AP / September 30, 2013, 11:38 AM

Syrian FM claims some rebels eat human hearts



[Open publisher's homepage...](#)

[Open this page...](#)

Most Popular

- 01 / "Breaking Bad" mania leads up to finale
 53955 views
- 02 / On the verge of a government shutdown: Now what?
 50999 views
- 03 / New Amanda Knox trial

Figure 12: iDiversiNews – News Item

Figure 12 shows an original news item surrounded by additional metadata. Once a news item is touched in the news story details or the ranked news items view, the original news items is displayed embedded in the view. In the top panel metadata extracted from the news item is displayed including a key text snippet, the publishing date and time, related people and organization as well as the categories classifying the news item. In the bottom panel a user can open the publisher's homepage or the specific news item and reload the news item or switch to the prior/next news item in the list.

The bottom part of the news story details view covers the "read news item" use case described in section 1.4.3 that extends the "browse ranked news" use case.

3 Demo Materials

Beside this textual deliverable, further materials for demonstration of the showcase are provided. On the one hand the actual showcase application, in particular the mobile frontend iDiversiNews. On the other hand additional dissemination materials feature the showcase, are provided to the public and may be used in live demonstrations.

3.1 Showcase Application

The system requirements and setup of the showcase applications are listed below.

3.1.1 System Requirements

The system requirements for end users to run and use the applications are limited to the devices required to run the frontend application, in particular Apple iPad running an updated version of iOS. Furthermore internet access is required to download the iDiversiNews app and subsequent for communication between app and backend services. As of the time of writing the acquisition of an official developer account by the responsible partner and the following review process of the app have not been finished. Information about the status of this process as well as contact details to request the current developer version are provided on the showcase website.

Services which are required by the iDiversiNews app are provided by the responsible partner. Sustainability and maintenance of these services is guaranteed for at least two years after the end of the project. The re-use of this technology in upcoming research projects will probably extend the availability of the services beyond the guaranteed timespan. The setup of the backend components is beyond the scope of this document.

3.1.2 Setup

The setup of the iDiversiNews app follows the default installation routines for apps via the App Store in iOS directly on the mobile device, e.g. the iPad. A link to the app website in the App Store will be provided on the showcase website after the deployment process is finished. A developer version may be requested in the meantime.

3.2 Dissemination Materials

Additional material featuring the showcase is provided on the showcase website, including the website itself. The showcase website mentioned in the following subsection is part of the project' website.

3.2.1 Website

This deliverable is aligned by a website to feature the showcase in the public. All information in relation to the showcase is collected at

<http://render-project.eu/showcase>

The showcase website provides background information, motivation and use cases as well as the links to the application, to demo websites of individual components and to dissemination material.

3.2.2 Screencast

The screencast demonstrates the textual and illustrated instructions for the web app given on the website as a step by step video tutorial with voice-over comments.

3.2.3 Flyer

An information flyer for the RENDER project was produced. Beyond other case studies the DiversiNews showcase was introduced regarding RENDER's mission.

3.2.4 Poster

A poster has created to visualize the main concepts and the architectural specifications. It is available for download at the showcase website.

4 Conclusions

The showcase of the RENDER project has been presented in this deliverable. A detailed motivation, including the scenario, potential stakeholder, chosen approach and derived use cases have been described. DiversiNews, the showcase application, showed how technology developed within the project can be combined in one application that tackles issues appearing in the world in a new unique way. A set of additional dissemination materials is provided along with the showcase.

It has been shown that the approach described in section 1.3 has been successfully implemented. In particular all the function areas mentioned in the approach are covered. Supported by DiversiNews described in section 2, a user is able to get central news access (cp. section 1.3.1), to access content aggregated and ranked (cp. section 1.3.2) according to his viewpoint defined by different aspects of diversity, which have been revealed by the system (cp. section 1.3.3). Finally iDiversiNews provides the user with an intuitive user interface that visualizes the results and provides an easy definition of the viewpoint (cp. section 1.3.4). All derived use cases defined in section 1.4 are tackled by the showcase application.

The scenario motivated the opportunities but also the issues of a facet in the digital world, the digital publishing of news. DiversiNews is step in the right direction, enabling interested users to cope with the vast amount of information that is accessible, while still being able to profit from the opportunities with an appropriate effort of time. The feasibility of the approach was shown by the evaluation of DiversiNews in respective deliverables.

Finally the showcase frontend app in combination with the backend services is on the way to production use, e.g. downloadable for everyone⁶ in a productive environment. It shows the practicability of the approach in terms of performance and scalability.

⁶ Currently everyone in terms of iPad owners

References

- [1] Jean-Yves Delort and Enrique Alfonseca. DualSum: a Topic-Model based approach for update summarization. Proceedings of ACL, 2012.
- [2] Mitja Trampuš and Blaž Novak. Internals of an aggregated web news feed. Proc. of SIKDD, 2012.
- [3] Tadej Štajner, Deli Rusu, Lorand Dali and Blaž Fortuna. Enrycher: service oriented text enrichment. Proc. Of SiKDD, 2009.
- [4] Tehranian, M. (2002). Peace Journalism Negotiating Global Media Ethics. The Harvard International Journal of Press/Politics, 7(2), 58-83.
- [5] Vīķe-Freiberga, V., Däubler-Gmelin, H., Hammersley, B. and Maduro, L. M. P. P. (2013). A free and pluralistic media to sustain European democracy, Brussels: High Level Group on Media Freedom and Pluralism.

Annex A DiversiNews Web App

A predecessor of the showcase application has been the DiversiNews web application⁷. DiversiNews has been described and evaluated in detail in the deliverables D5.2.2 (Diversified News Service), D5.2.3 (Refined Diversified News Service) and D5.2.4 (Evaluation of the diversified news service). Feedback and the evaluation results have been incorporated in the development of the showcase application.

In contrast to the showcase app the web application was not intended for productive use. Nevertheless, it already provides all the major functionality described in this deliverable for the showcase app, but running disconnected from the live news item stream on a static data set. The intermediate and final evaluations in D5.2.3 and D5.2.4 have been conducted on the prototype. For the sake of completeness and to show the development after the final evaluation, screenshots of the web application for each of the user interface functionalities of iDiversiNews described in section 2.3.4 are – if supported – shown in the following.

⁷ <http://aidemo.ijs.si/diversinews/>

DiversiNews

A tool for interactive exploration of news.

... or choose a recent news cluster:

- [Microsoft Releases Tweaks, Performance Update For Surface](#) 2012-11-15
- [Expect BlackBerry 10 Devices to Start Shipping in February](#) 2012-11-14
- [Caffeinated Drink Cited in Reports of 13 Deaths](#) 2012-11-15
- [Managing Diabetes Doesn't Have To Be Difficult](#) 2012-11-13
- [UPDATE 1-BP expected to admit to criminal misconduct in 2010 spill](#) 2012-11-15
- [Germany, France eke out third quarter growth](#) 2012-11-15
- [5 Reasons to Believe Carmelo Anthony Is Ready to Finally Evolve](#) 2012-11-15
- [Grand Theft Auto 5's trailer two ensures you have one good day](#) 2012-11-14
- [BioWare responds to Black Ops 2 mix-up with Mass Effect Trilogy giveaway](#) 2012-11-14
- [Toyota Issues Another Massive Recall](#) 2012-11-14
- [College Football 2012 Schedule: Stanford visits Oregon in key Pac-12 battle](#) 2012-11-14
- [Japan set for December elections](#) 2012-11-14
- [MICHIGAN HOCKEY: Wolverines fall to Fighting Irish](#) 2012-11-17
- [\\$1.5 Million Worth Of iPads Stolen From JFK! \(AAPL\)](#) 2012-11-15
- [Facebook stock costs will wipe out profit](#) 2012-11-14
- [Report: Jets favor Mark Sanchez](#) 2012-11-14

Figure 13: DiversiNews – News Stories (cp. Figure 9)

DiversiNews

A tool for interactive exploration of news.

«« ... or return to the cluster listing

Summary:
 Choose summarization algorithm: Type1 (current) Type2

"Although this is a smaller rise than recorded in recent months, and there are good reasons to expect long-term unemployment to stop rising in coming months, there are now almost 900,000 people in the UK who have been out of work for more than a year.

That's why we've committed to supporting the hardest-to-help people over a two-year period through the Work Programme so that we can help them overcome their barriers to work and get them into sustainable jobs."

The figures were published on the same day that the Bank of England downgraded its growth forecast for 2013 to around 1% and warned that output would remain below pre-financial crisis levels for the next three years.

The World Failed Sri Lanka. And Continues to Do So.
 The top story buzzing around the UN today the soon-to-be-released report on the failure to protect civilians caught in the final days of the Sri Lankan civil war. Parts of the report were already leaked to the BBC, and Ban Ki Moon is expected to make it p...
[undispatch.com](#) (105 69775812 eng -0.638 [0.0,0.0])

Fiscal Cliff? Obama Urged Not To Panic
 WASHINGTON -- Republicans and deficit hawks are raising unnecessary alarm over the so-called "fiscal cliff" to pressure President Barack Obama into a "grand bargain" he shouldn't make, progressive economists and scholars said Tues...
[huffingtonpost.com](#) (123 69783042 eng -0.000 [37.2,-95.8])

Petraeus scandal creates 'nightmare' for Mount Laurel lawyer and family
 "I've had no sleep, to be honest with you," Khawam, 40, said at his Camden County office on Tuesday. Before discussing his shift from the quiet life of a suburban New Jersey attorney to the object of inquiries from everyone "from CNN ...
[phillyburbs.com](#) (179 69807801 eng +0.075 [38.0,-97.0])

Refine the search results

Focus on news about ?

Focus on news coming from ?

☐ enable

Focus on news with sentiment that is ?

negative positive

Figure 14: DiversiNews – News Story Details/Ranked News Items (cp. Figure 10 and. Figure 11)

The New York Times True Colors iPhone app iPad app Android phone app Android tablet app More

Log in Create Account

September 13, 2013

HUFF POST MEDIA

Edition: U.S. Search The Huffington Post

Like 69k Follow

FRONT PAGE POLITICS BUSINESS ENTERTAINMENT TECHNOLOGY TV ARTS BOOKS LIVE ALL SECTIONS

Media Eat The Press • Calderone: The Backstory • Third Metric

Katie Couric May Replace Barbara Walters On 'The View'

WATCH: Christiane Amanpour Angrily Calls For Syria Intervention

Paul Krugman Vs. Twitter

THE BLOG Featuring fresh takes and real-time analysis from HuffPost's signature lineup of contributors

HOT ON THE BLOG Indra Nooyi Josh Silver Sen. Bernie Sanders Albert Brooks

Eric Dezenhall
CEO, Dezenhall Resources

GET UPDATES FROM ERIC DEZENHALL
FAN RSS EMAIL Like 1

Why Managing Scandals Is Different Now

Posted: 11/16/2012 3:09 pm

React > Amazing Inspiring Funny Scary Hot Crazy Important Weird

Read more > Damage Control, David Petraeus, PAULA BROADWELL, Crisis Management, Scandal, Media News

SHARE THIS STORY

3 9 1 0

Share Tweet Email +1

Submit this story

Whenever a mess like the General David Petraeus affair blows up, I am asked by friends, acquaintances, students and journalists to handicap the state of play. The challenge is that no matter how hard I try, I'm lucky if I can get across one semi-cogent sound bite or bottom-line conclusion in most exchanges.

What follows, then, are some observations you won't hear on cable TV about how crisis

MOST POPULAR

23 Things Every Woman Should Stop Doing

Dad Wears Embarrassing Outfit To Teach His Daughter A Lesson

Figure 15: DiversiNews – News item (cp. Figure 12)

Annex B SiKDD 2013 Submission

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

Mitja Trampuš, Flavio Fuat, 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 project¹ 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 application² 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 areas⁵ and possible disease outbreaks⁶[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 listings 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

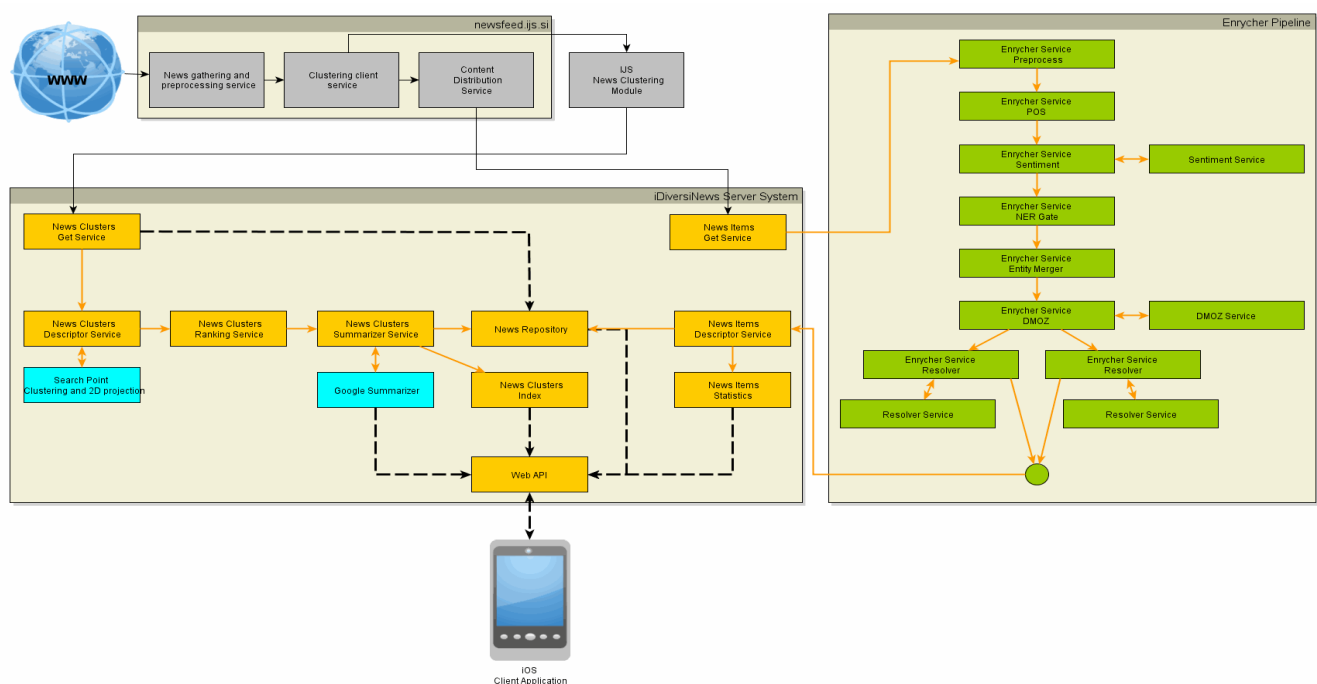


Figure 1: iDiversiNews System Architecture, in part shared with DiversiNews

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).

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

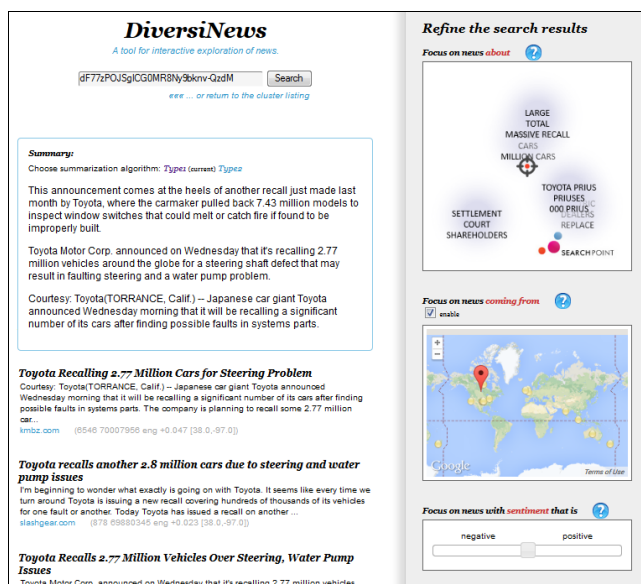


Figure 2. DiversiNews main interface.

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

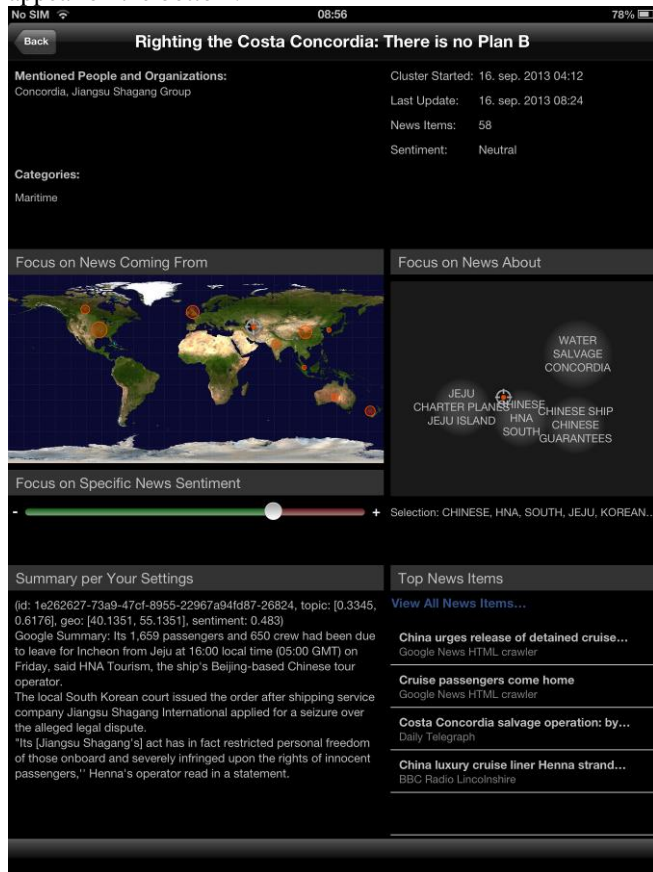


Figure 3: iDiversiNews

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 both quantitatively, from 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 Fuat, Erik van der Goot (2012). Media monitoring of public health threats with MediSys. *Living in Surveillance Societies*.
- [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.