Context-Aware User-Driven News Recommendation

Jon Espen Ingvaldsen
Norwegian University of Science and Technology, Department of Computer and Information Science, Trondheim Norway
jonespi@idi.ntnu.no

Özlem Özgöbek
Department of Computer Engineering, Balikesir University, Balikesir, Turkey
ozlem.ozgobek@balikesir.edu.tr

Jon Atle Gulla
Norwegian University of Science and Technology, Department of Computer and Information Science, Trondheim Norway
jag@idi.ntnu.no

ABSTRACT
Recommender systems match available contents with users' contexts and interests. With linked data knowledge bases we can build recommender systems where user interests, their context and available contents are modeled in terms of real world entities. In this demo paper we will describe existing academic news recommender systems and the Smartmedia prototype in particular. This prototype shows how we can combine available technologies like semantics, natural language processing and information retrieval to construct personalized and location aware recommendations on a continuous stream of news information.

Categories and Subject Descriptors
H.4.7 [**Information Systems Applications**] Communications Applications – *Information browsers*

General Terms
Algorithms, Design, Experimentation, Human Factors.

Keywords
Recommender system, news, mobile, natural language processing, named entity disambiguation.

1. INTRODUCTION
News organizations and libraries catalog their contents. These catalogs are traditionally constructed using controlled vocabularies with limited context information about what the catalog identifiers really mean. Even though a news article can be annotated with *"Barack Obama"*, there is no catalog information data saying he is the president of USA. The same article might be annotated with *"Boise"*, but we do not have data saying that it is a city and its longitude – latitude coordinates. Such extensive meta data attributes are valuable when we want to expose and personalize archive contents to a context aware user setting.

With large knowledge bases, such as WikiData and Yago, we get access to extensive databases of real world entities that are precisely described and structurally interlinked. By indexing news archives on such entity representations we can build news recommender systems that allow us to construct more ambitious catalog queries. For instance, we can retrieve all news articles from January 2015 related to the president of USA and locations within the range of 50 miles for the city center of Boise. This entity representation provides better solutions to challenges of news recommender systems like unstructured content, serendipty and synonmy [9].

The Smartmedia project[1] at NTNU targets construction of context aware news experiences based on deep understanding of text in continuous news streams [6, 11]. The goal of the Smartmedia project is to deliver a mobile and context aware news experience based on deep understanding of textual contents, combining both geo spatial exploration and context aware recommendations. In this project, we have implemented a prototype of a news recommender system where news articles are processed and represented in terms of WikiData entities. In this demo paper we will describe this news recommender prototype, its stream based data processing pipeline and its context-aware recommendation features.

Section 2 describe related work, followed by a description of the Smartmedia prototype. Section 3 describes how its stream based data processing pipeline is constructed and Section 4 presents its mobile user interface and interaction principles. Conclusions and future work is given in Section 5.

2. RELATED WORK
The prototype system, described in this paper, share similarities to other academic news applications such as NewsStand [10, 12] and News@Hand [2, 3]. Both these systems map textual news contents to entities defined in a knowledge base. NewsStand targets geo spatial exploration of news. It is an example application of a general framework developed to enable people to search for information using a map query interface. It utilizes maps both to explore and find news stories and to visualize and present single news events. News@hand combines textual features and collaborative information to make news recommendations. It uses Semantic Web technologies to describe the news contents and user preferences. Both news items and user profiles are represented in terms of concepts appearing in domain ontologies, and semantic relations among those concepts are exploited to enrich the above representations, and enhance recommendations. Both NewsStand and News@Hand have user interfaces targeting desktops and larger device screens. They both provide user control over the retrieved set of news, either through a map or category based navigation or preferences settings. The Smartmedia prototype targets mobile devices and combine some of the geospatial and ontological news data representation features found in both of two other system.

Asikin and Wörndl [1] propose news recommendation techniques based on a location aware context model. The spatial model proposed in this work contains geographical information like latitude and longitude, and a human described physical character of a location and place identity, which represents the location's meaning and significance. In this work it is also focused on the improvement of serendipity problem.

Building user profiles is an important aspect of recommender systems. Meguebli and Kacimi [8] propose an approach to build user profiles based on the comments added to news articles. To do that, entities are extracted from the user comments. Then the related aspects are extracted from the news articles. By using all the articles that a user read, a user profile is created based on the extracted aspects.

Another personalization approach in news recommender systems is Hermes framework [7]. In this work, in addition to the personalization an ontology based approach is used to recommend news articles. Also in [5] a semantic news recommendation framework which is called Athena is proposed. In this work, CF-IDF (Concept Frequency - Inverse Document Frequency) method is proposed which is the application of TF-IDF (Term Frequency - Inverse Document Frequency) method to semantic recommenders. CF-IDF is a selective method compared to TF-IDF where it considers only the key concepts in the news articles where TF-IDF considers all terms.

[1] http://research.idi.ntnu.no/SmartMedia

INRA 2015 in conjunction with RecSys '15, September 16-20, 2015, Vienna, Austria

3. IMPLEMENTATION

The backend of the news recommender prototype developed is constructed as a pipeline of operations transforming Rich Site Summary (RSS) entries and raw text data into a semantic and searchable representation. The pipeline and its operations are implemented with using the Apache Storm[2] framework. This distributed computing framework enables scalability and ability to handle large amounts of news items from a magnitude of publishers continuously.

There are five steps involved in the data processing. The first step creates an input stream by continuously monitoring a set of RSS feeds from a wide range of news publishers. Whenever a new news item occurs, RSS entry properties such as the title, lead text and HTML sources are retrieved. The HTML sources are parsed and cleaned to extract a representative body text. In the second step, natural language processing operations such as language identification, sentence detection and part-of-speech tagging is applied to extract entity mentions from the textual data. The third step uses supervised models to map entity mentions to referent entities in the WikiData knowledge bases. These models combine textual similarities, WikiData graph relations and entity frequencies and co-occurrence statistics to classify the relevance of multiple referent candidates. First Story Detection is applied in the fourth step to group news items describing the same news story. In the fifth step this semantic representation is indexed and made searchable. As this backend architecture is stream based, it is able to index and promote recent news items soon after they are discovered.

WikiData is the community-created knowledge base of Wikipedia[13]. Since its public launch in 2012, the knowledge base has gathered more than 15 millions entities, including more than 34 million statements and over 80 million labels and descriptions in more than 350 languages[4]. Most geographical entities in WikiData provide a reference to Geonames containing more detailed geographical properties. In the implementation of the Smartmedia prototype, the entity information from these knowledge bases where indexed in a Lucene[3] based search index. This index makes the entities searchable and creates a foundation for addressing entity labels, descriptions and aliases, entity relations and geospatial properties.

Figure 1 shows an example of a news article from the Guardian where the text is parsed and enriched with WikiData entity annotations. The fields and nested data structure in this figure are similar to how the news stories are stored and indexed in the Lucene based index. By running the news text from the news article in the figure through the data processing pipeline, we identified nine WikiData entities, including Bedfordshire, Home Office and Theresa May. Note that the news texts and list of entities and associations in the figure is shortened. All entities contain a textual description and a list of associations. These associations are typed relations to other WikiData entities. We can see that Bedfordshire contains eight such entity associations. Examples of entities linked and related to Bedfordshire are the *instance of* relations to *Ceremonial county of England* and *Administrative territorial entity of the United Kingdom*. Both Bedfordshire and Home Office are additionally described with geospatial properties. In this case the geospatial properties are longitude – latitude pairs, but the implementation allows for any geo spatial shape decribed as valid Geojson[4].

When a user is opening the news app on the mobile a request containing user id, location and preferences are sent to the backend. Here, a multi factor search query is formed to retrieve relevant news entries from the index.

4. USER INTERFACE

A web-based and responsive user interface is developed to make the news stream contents explorable on mobile devices. In this interface, the user is allowed to extract news items that are relevant to the geo special locality context, personal interests and given point of time. These three relevance factors are customizable and the user can select whether or not they should influence the retrieved news items.

To customize the geographical locality, the user specifies a circular relevance region on a map. Figure 2a shows an example of such a relevance region. By default, the relevance region is set to users current GPS location with a 50 km radius. By moving the region or modifying the radius, users can generate a local newspaper for any region of the world. If the location factor is disabled, it means that the system is recommending news from any location in the world and news that are not containing location information.

In the current Smartmedia prototype, we have predefined a handful of user interest profiles. Each user profile contains an alias and a weighted vector of WikiData entities. Examples of predefined profiles in the system are stock trader, soccer fan, technology geek, etc. By selecting any of these interest profiles, the retrieved news will be influenced and biased towards the interest topics. When the personal interest factor is disabled, the user retrieve a news composition which is general and without such bias.

By changing the time-factor, the user is presented with a calendar where can move in time and retrieve either recent or historic news items. When, the time-factor is disabled the user will retrieve news solely based on the other relevance factors (location and personal interests).

Figure 2b shows an example of how news stories are presented. Here we see the same article as we had in Figure 1. The three circular buttons on the bottom of the screen allow users to toggle whether their locality, personal interest profile and time setting such influence news story retrieval.

By clicking on a news story, the user gets the ingress of the news story and a list of the most salient entities for the selected news story. Figure 2c shows the ingress and relevant WikiData entities from the news article about Theresa May. As we can see, our news story about politics and terror related to Syria, Theresa May, ISIL and Sky News. By hovering these items, the user is presented with their textual WikiData description. On figure 1c, we can see that the WikiData entity for Theresa May contains the description *"British politician"*.

In general, the three buttons at the bottom of the screen for location, interest profile and time can at any time be activated and de-activated in combinations to provide very different recommendation strategies. For example, keeping all buttons active with default parameters means that the system will recommend news articles that have recently takes place in the vicinity of the reader and are consistent with her profile. A screencast video describing the features of the system and its user interface is available at https://vimeo.com/121835936

5. CONCLUSIONS AND FUTURE WORK

Many see the full stack of semantic web technologies as a complex implementation of some really simple and good ideas about adding meaning to data. There are great rewards in understanding the full stack and what it can do, but most news organizations find great rewards by looking into linked data in combination with traditional information retrieval techniques.

In this paper we have shown a prototype of a news recommender system that demonstrates some of the context and geo spatial aware features online news services can achieve by using available and open knowledge bases and data processing and storage technologies.

Future work for the Smartmedia prototype will focus on improvement on entity linking qualities and evaluations of user needs. The user evaluations will look into to which extent users find the ability to control their news feed in terms of location, interest profile and time valuable and useful.

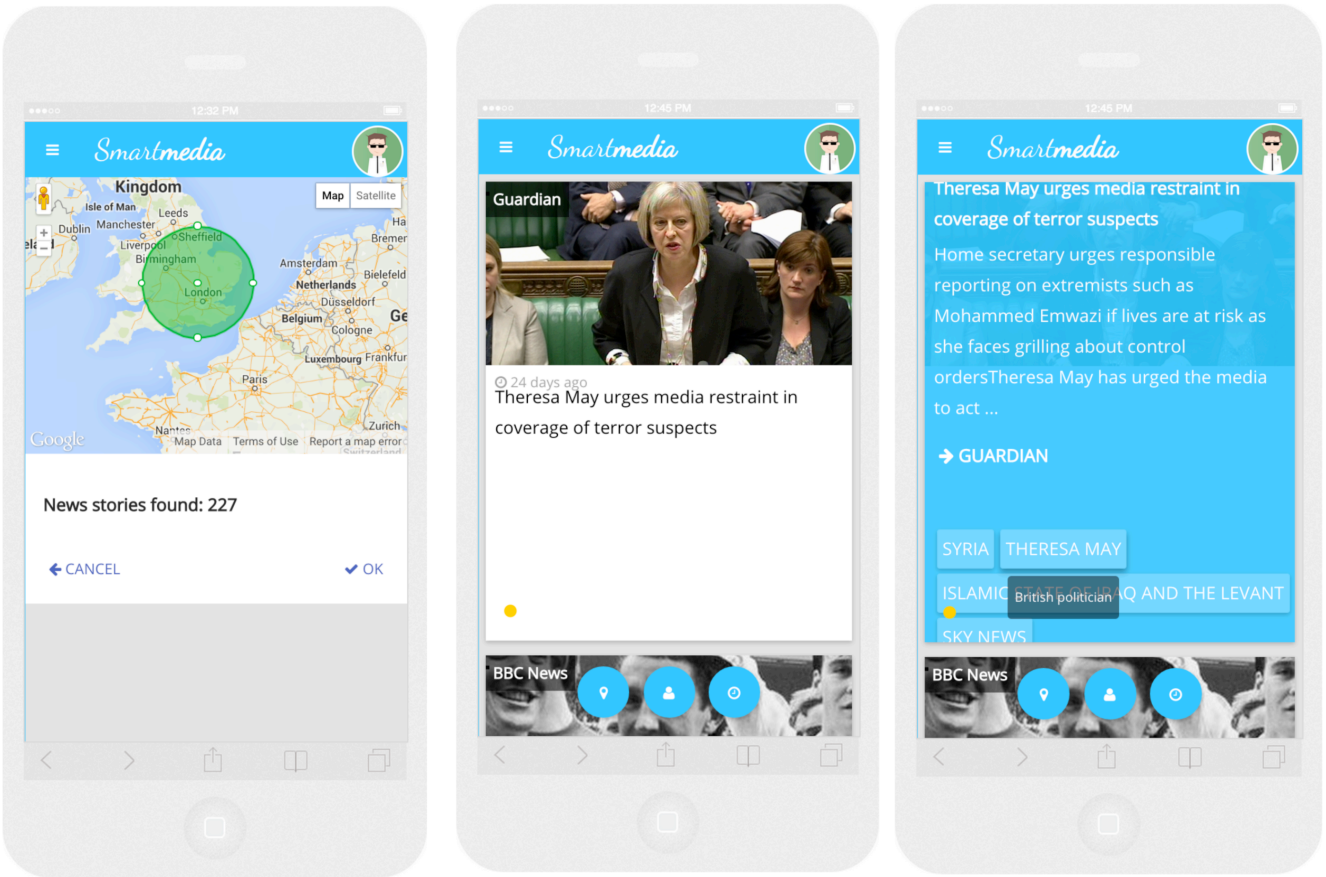
[2] http://storm.apache.org/

[3] https://lucene.apache.org/core/

[4] http://geojson.org/

```
articleId: "Guardian_254439378"
type: "article"
title: "Theresa May 'allowed state-sanctioned abuse of women' at Yarl's Wood"
leadText: "Shadow home secretary criticises minister after TV documentary alleges rape and self-harm at detention centre were
ignoredTheresa May, the home secretary, has been accused of allowing the "state-sponsored abuse of women" at the Yarl's Wood detention
centre after a Channel 4 investigation uncovered guards ignoring self-harm and referring to inmates in racist terms.Yvette Cooper..."
entities:  [ 9]
          0:   {
                    entityId: "Q23143"
                    name: "Bedfordshire"
                    description: "county in England"
                    associations:  [ ... 8]
                    shape:  {
                              type: "Point"
                              coordinates:  [ 2]
                                        0:  -0.41666666666667
                                        1:  52.083333333333
                    }
          }
          1:   {
                    entityId: "Q763388"
                    name: "Home Office"
                    description: "ministerial department of the Government of the United Kingdom"
                    associations:  [ ... 3]
                    shape:  {
                              type: "Point"
                              coordinates:  [ 2]
                                        0:  -0.129948
                                        1:  51.4958
                    }
          }
          2:   {
                    entityId: "Q264766"
                    name: "Theresa May"
                    description: "British politician"
                    associations:  [ ... 21]
          }
```


Figure 1. Example of a news article enriched with WikiData entities.



a) b) c)

Figure 2. Screenshots from the Smartmedia prototype. a) The map query interface. b) Presentation of news stories. c) Presentation of news details.

6. REFERENCES

[1] Asikin, Y. and Wörndl, W. 2014. Stories around You: Location-based Serendipitous Recommendation of News Articles. *Proceedings of 2nd International Workshop on News Recommendation and Analytics*. (2014).

[2] Cantador, I., Bellogín, A. and Castells, P. 2008. News@ hand: A semantic web approach to recommending news. *Adaptive hypermedia and adaptive web-based systems*. (2008).

[3] Cantador, I., Bellogín, A. and Castells, P. 2008. Ontology-based personalised and context-aware recommendations of news items. *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 1, (2008).

[4] Erxleben, F., Günther, M. and Krötzsch, M. 2014. Introducing Wikidata to the Linked Data Web. *The Semantic Web–ISWC 2014*. (2014).

[5] Goossen, F. and IJntema, W. 2011. News personalization using the CF-IDF semantic recommender. *Proceedings of the International Conference on Web Intelligence, Mining and Semantics (WIMS)*. (2011).

[6] Gulla, J.A., Ingvaldsen, J.E., Fidjestøl, A.D., Nilsen, J.E., Haugen, K.R. and Su, X. 2013. Learning User Profiles in Mobile News Recommendation. *Journal of Print and Media Technology Research*. II, 3 (2013), 183–194.

[7] IJntema, W. and Goossen, F. 2010. Ontology-based news recommendation. *Proceedings of the 2010 EDBT/ICDT Workshops*. (2010).

[8] Meguebli, Y. and Kacimi, M. 2014. Building rich user profiles for personalized news recommendation. *Proceedings of 2nd International Workshop on News Recommendation and Analytics*. (2014).

[9] Ozgobek, O., Gulla, J. and Erdur, R. 2014. A survey on challenges and methods in news recommendation. *In Proceedings of the 10th International Conference on Web Information System and Technologies (WEBIST 2014)*. (2014).

[10] Samet, H., Sankaranarayanan, J., Lieberman, M.D., Adelfio, M.D., Fruin, B.C., Lotkowski, J.M., Panozzo, D., Sperling, J. and Teitler, B.E. 2014. Reading news with maps by exploiting spatial synonyms. *Communications of the ACM*. 57, 10 (Sep. 2014), 64–77.

[11] Tavakolifard, M., Gulla, J.A., Almeroth, K.C., Ingvaldesn, J.E., Nygreen, G. and Berg, E. 2013. Tailored news in the palm of your hand: a multi-perspective transparent approach to news recommendation. *WWW '13 Companion Proceedings of the 22nd International Conference on World Wide Web*. (May 2013), 305–308.

[12] Teitler, B. and Lieberman, M. 2008. NewsStand: A new view on news. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. (2008).

[13] Vrandečić, D. and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*. (2014).