DaCENA: Serendipitous News Reading with Data Contexts

Matteo Palmonari[1](✉), Giorgio Uboldi[2], Marco Cremaschi[1], Daniele Ciminieri[2], and Federico Bianchi[1]

[1] Università degli Studi di Milano-Bicocca, Milano, Italy
{palmonari,cremaschi,f.bianchi}@disco.unimib.it
[2] Politecnico di Milano, Milano, Italy
{giorgio.uboldi,daniele.ciminieri}@gmail.com

Abstract. DaCENA (Data Context for News Articles) is a web application that showcases a new approach to reading online news articles with the support of a data context built from interlinked facts available on the Web of Data. Given a source article, a set of facts that are estimated to be more interesting for the readers are extracted from the Web and presented using tailored information visualization methods and an interactive user interface. By looking at this background factual knowledge, the reader is supported in the interpretation of the news content and is suggested connections to related topics that he/she can further explore.

1 Introduction

In this paper we present DaCENA (Data Context for News Articles), a web application that let a user read a news article while contextually exploring additional content extracted from a semantic Knowledge Base (KB). Like in other annotated corpora, a set of named entities of a KB is extracted from text using an entity linking tool. The key feature of DaCENA is the approach to model, organize, visualize and interact with the additional content presented to the reader: at the side of an article, we present several *semantic associations* [1], i.e., paths of connected facts extracted from a linked data source. The associations connect a main entity representing the article's topic with many other named entities found in the text. The set of extracted associations defines the *Data Context of the article*. For example, an article entitled “A Threat to Spanish Democracy” published in the New York Times on 7th, November 2014, discusses the threat of Catalan nationalism for Spanish democracy. *Catalonia* is deemed to be the main entity of the article, while other thirteen entities are found, including *Catalan nationalism*. The Data Context consists of 278 associations, among which we find that *Catalonia is birth place of Artur Mas I Gavarro, who is the president of Convergence and Union* (a party)*, which has Catalan nationalism as ideology.*

The Data Context can be very large, e.g., up to dozens of thousands associations, and contain many uninteresting associations, e.g., *Barack Obama is born in America.* To rank and filter out the associations based on their expected

F. Gandon et al. (Eds.): ESWC 2015, LNCS 9341, pp. 133–137, 2015.
DOI: 10.1007/978-3-319-25639-9_26

interest for the reader, we evaluate every association using a novel *serendipity* measure. This measure considers the relevance of the association in relation to the news article and the unexpectedness of the association in the KB. The balance between relevance and serendipity can be determined by the user at runtime according to his/her preferences. A set of top-k serendipitous associations is presented to the reader, while more associations can be visualized on demand. DaCENA can be tested online on several articles extracted from the New York Times at the address http://dacena.densitydesign.org/.

DaCENA is targeted to two different kinds of users: readers who are interested in data-supported stories, in the vein of data journalism [3], and journalists in the newsroom. A reader can leverage the explored data to better understand the context of the story told in the article. For example, from the above mentioned association between *Catalonia* and *Catalan nationalism*, the reader discovers the name of a prominent politician (currently leading the Catalonia government) and of his party, which were not mentioned in the article. In addition he/she learns that such a party supports a nationalist vision. A journalist may use DaCENA to have inspiration for a new story. For example, in other associations found by DaCENA, e.g. the ones between *Catalonia* and *Autonomous countries of Spain*, he/she can find out about other autonomous communities. These findings may inspire a new story about the political landscape behind nationalist ideologies in Spain and a comparative analysis of different political movements in favour of autonomy. Ultimately, DaCENA aims to be a first steps towards data journalism based on the analysis of relational data.

The novelties introduced in DaCENA can be summarised as follows. (1) Data journalism initiatives has focused primarily on statistical data analysis and visualization [3]. Semantic technologies have been applied in the news domain mainly to connect articles based on the co-occurrences of named entities [7], or to extract relations among entities from large news corpora (e.g., in the NewsReader project[1]). We present a new application domain for semantic technologies, which could be defined as *relational data-journalism* because of its focus on relational data. (2) We introduce a measure to evaluate the serendipity of semantic associations in relation to an input text. (3) We introduce a novel interactive visualization interface, which combines text reading and data exploration.

2 DaCENA

The DaCENA framework consists of two main components. Text & Data Analyzer is a component responsible for processing and storing the information used to build the data context of an article. Contextual Explorer is an interactive user interface that let the user read the articles enriched with data contexts. The components exchange data in the JSON format via HTTP APIs. Processing the articles may require significant amount of time (up to thirty minutes) if

[1] http://www.newsreader-project.eu/.

semantic data are fetched by querying a SPARQL endpoint as we currently do[2]. Therefore, texts and data are processed off-line so as to make the interactive visualization features as much fluid as possible. The two components and the main features of DaCENA are explained here below.

Text and Data Analyzer. The data context extraction process is applied to any article scraped from a news source and consists of the following steps.

Entity linking and main topic extraction. The whole text of an article is passed to the DBpedia Spotlight service, which recognizes named entities in the text and links them to DBpedia. The confidence used by Spotlight to tune the precision vs recall of the algorithm is dynamically set based on the number of retrieved entities. When too many or too few entities are found, we dynamically adjust the confidence threshold (we use confidence values in the range $[0.20, 0.35]$). One of the entities recognized in the text, the one that is mentioned more frequently, is selected as main topic of the article.

Data Context Extraction. Given the main topic m, i.e., the main entity, we find every possible semantic associations between m and every other extracted entity e, which have a length equal or shorter than a given threshold (set to 3 in the current demo). A semantic association between two entities is defined as a semipath connecting the two entities, such that every other node in the path is also an entity. By considering only links between entities (we do not consider concepts or literals) we focus on sequences of relational facts. We traverse the graph in any possible directions when we search for the associations because the incoming links of an entity may represent unexpected connections, which are deemed valuable for the readers (e.g., *Catalonia is the birth place of Artur Mas I Gavarro*, in the example discussed in Sect. 1). The Data Context Graph of an article is the set of all the associations extracted from it. The Data Context Graph is retrieved by constructing and submitting a set of queries to the DBpedia SPARQL endpoint. Every extracted association is indexed and stored in a relational database to speed-up data fetching at runtime.

Data Context Evaluation. We evaluate each association of the Data Context Graph with measures that return values in the range $[0, 1]$, to ease their composition. The objective of DaCENA is to present serendipitous associations to readers. By referring to a previous definition [9], we can define serendipity as unexpected relevance of a finding. We propose to capture this definition by modeling *serendipity* as a linear combination of two measures: *relevance*, which estimates the degree of match between the topics covered by the article and the semantic associations, and *unexpectedness*, which estimates the level of surprise that associations can generate in the reader. We define a parametric serendipity measure applied to an article a and an association π as $s(\pi, a) = \alpha * r(\pi, a) + (1 - \alpha) * u(\pi)$, where $r(\pi, a)$ is the relevance of π wrt the article a, $u(\pi)$ is the unexpectedness

[2] In our experiments with the DBpedia SPARQL endpoint, we observed that the time needed to process an article time is largely unpredictable and not correlated to the size of the data context.

of π, and α is a parameter in the interval $[0, 1]$ that can be used to adjust the weight assigned to relevance and unexpectedness; for example, for $\alpha = 0.3$ the unexpectedness component of serendipity is emphasized, while for $\alpha = 0.7$ the relevance component of serendipity is emphasised. To capture unexpectedness we use a measure proposed in previous work [1], named Rarity, which evaluates how scarce the use of a property is in a dataset. In this way, rarity rewards those associations that use properties that are uncommon in the KB. Because we could not find a definition of *relevance of a semantic association wrt to a reference text*, we defined our own relevance measure inspired by Information Retrieval [5]. For each association (π) we concatenate the abstracts of each entity occurring in π so as to form a virtual document d^{π}, which is represented by a vector of weighted terms. Each term is weighted using TF-IDF [5], where IDF is computed over the collection of all the abstracts collected for an individual article. Analogously, we build a virtual document d^{a} for the article a. We then compute the relevance of π wrt a as the cosine similarity between the two documents d^{π} and d^{a}. Finally, before being combined into the serendipity measure, relevance and unexpectedness are normalised in the range $[0, 1]$ using the min-max normalisation method, which preserves the relative distances between values [10]. We use Elastic Search[3] to process, index and match the virtual documents. The computation of serendipity s is very efficient if relevance r and unexpectedness u are available: r and u are computed for every association off-line, while serendipity is dynamically computed at run-time by letting users specify the parameter α, which determines the balance between relevance and unexpectedness. Serendipity is used to rank the associations in the Data Context Graph and select a set of top-k associations shown to the reader.

To the best of our knowledge, serendipity and relevance, defined for a semantic association in relation to an input text, are original contributions of our work, while unexpectedness has been taken from previous work [1]. The few computational definitions of serendipity that have been proposed, e.g., see [6,9], are not easily applicable to linked data because defined in very different contexts. Graph-based measures proposed to evaluate the relatedness between entities in semantic graphs [2,4] could be useful to enrich the definition of relevance. However, these measures do not consider a reference text as our relevance measure does, which is a key asset in our application scenario.

Contextual Explorer. The user interface of DaCena has been designed to offer the user an interactive environment to read the news article and visually explore the semantic associations simultaneously. The aim is to offer the user an innovative reading experience based on an exploration process characterized by the ?overview first, zoom and filter, details on demand? browsing model typical of information visualization interfaces [8]. The interface presents on the left side the news article formatted in order to guarantee a good readability of the text. The named entities found are highlighted in yellow in the article as the main entity, which is presented at the beginning of the section. On the right side we can find the interactive graph generated by the correlations between the main

[3] https://www.elastic.co/products/elasticsearch.

entity (the big yellow node), the other entities mentioned in the article (the small yellow nodes) and the entities that occur in the associations but not in the article (the grey nodes). Clicking on the entities, both through the visualization and the text on the left, the user can filter the network and explore in detail all the paths from the main entity to the selected one. In the upper part of the interface the user can access different parameters to filter or expand the graph.

3 The Demo

The demonstration showcases the concept and novel data exploration interface of DaCENA. Users can read several articles from within the interactive interface and personalize the data context view. They can dynamically change the number of displayed associations, filter out associations based on their length, and look associations between the main topic and one entity of interests in more details. In addition, users can dynamically tune serendipity to favor unexpectedness or relevance; their preferences will result in a quick adjustment of the shown graph.

Despite we developed DaCENA for the journalism domain, the main ideas presented in this paper can be virtually applied to any domain where it is useful to explore a (relational) data context related to a text of interest (for example, in the forensic or in the music domain). By exploring linked data with DaCENA in the context of reading tasks, we also experienced some limits in DBpedia, when this KB is challenged to provide interesting data for end-users. We believe that DaCENA can stimulate interesting discussions on topics such as data quality, data value, and content specialization in socio-political and economic domains.

References

1. Aleman-meza, B., Halaschek-wiener, C., Arpinar, I.B., Ramakrishnan, C., Sheth, A.: Ranking complex relationships on the semantic web. IEEE Internet Comput. **9**, 37–44 (2005)
2. De Vocht, L., Coppens, S., Verborgh, R., Vander Sande, M., Mannens, E., Van de Walle, R.: Discovering meaningful connections between resources in the web of data. In: 6th Workshop on Linked Data on the Web, p. 8 (2013)
3. Gray, J., Chambers, L., Bounegru, L. (eds.): The Data Journalism Handbook. O'Reilly, Sebastopol (2012)
4. Leal, J.P.: Using proximity to compute semantic relatedness in RDF graphs. Comput. Sci. Inf. Syst. **10**(4), 1727–1746 (2013)
5. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
6. Noda, Y., Kiyota, Y., Nakagawa, H.: Discovering serendipitous information from wikipedia by using its network structure. In: ICWSM (2010)
7. Raimond, Y., Ferne, T., Smethurst, M., Adams, G.: The BBC world service archive prototype. J. Web Sem. **27**, 2–9 (2014)
8. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: VL, pp. 336–343. IEEE Computer Society (1996)
9. Sun, T., Zhang, M., Mei, Q.: Unexpected relevance: an empirical study of serendipity in retweets. In: ICWSM (2013)
10. Tarique Ahmad, P.S.K.S.M., Haque, S.: Privacy preserving in data mining by normalization. Int. J. Comput. Appl. **95**(6), 14–18 (2014)