Exploring History through Newspaper Archives

Jasna Škrbec, Marko Grobelnik, Blaž Fortuna
Artificial Intelligence Laboratory, Jožef Stefan Institute, Ljubljana, Slovenia

{jasna.skrbec,marko.grobelnik,blaz.fortuna}@ijs.si

Abstract. This demo presents a web application which implements a pipeline for searching and browsing through newspaper archives. It uses a combination of information extraction, enrichment and visualization algorithms to help users to grasp a large amount of articles normally collected in archives. Illustrative results show appropriateness of the proposed pipeline for searching and browsing news archives.

Keywords: newspaper archives, data mining, visualization

1 Introduction

Newspapers with a long tradition have gathered large news archives. In recent years some newspapers invested in the digitalization of archives, resulting in a million document corpora. The articles would typically be annotated with the metadata quality and quantity largely depending on the publishers and an archive type.

Typical search and browse interfaces do not work well with the archives. They are not specialized for news archives and as such do not take advantage of their inherent structure. Archives are not just a collection of articles, but they hide stories that can be presented in one or more articles, which happened in one or more locations during some period of time. News archives store rich historic information and can be turned into a valuable knowledge resource with a proper use of semantic technologies and data mining techniques. This demo introduces Archive Explorer, a system for annotating and presenting the archives in order to provide them an easier access to information and content connected with it.

The demo interface is designed to cover two common scenarios. The first scenario is the visualization of a particular article in the context of the overall archive. The second scenario is the summarization of a large collection of articles, providing glimpse of the events, entities, and topics covered by them. Interactive faceted search is used throughout the system to help with the navigation.

The system uses an annotation and contextualization pipeline, used to pre-process individual articles. In the annotation step, topics and entities occurring in the articles are recognized and linked to their corresponding resources in Linked Open Data (LOD) datasets. In the contextualization step, the articles are connected between each other based on their topicality, time, locations, events, important people, etc.

adfa, p. 1, 2011.

Connecting entities with LOD datasets brings additional relationships between entities into the archive. Disambiguation of an entity with its corresponding resource in DBpedia or Freebase can be used to link articles talking about the same entity, but using different labels. For example, if someone would search for Princess Diana, they would not find articles where she is mentioned with her real name Diana Spencer or with name that people gave her, Lady Di. But because the archive disambiguated with resources from LOD, the user would find all relevant articles, even if the query string does not directly occur in the article. Providing users with a consolidated view of the information that is crumbled across different articles is one of the core components of Archive Explorer and this heavily relies on LOD resources.

Linking entities with LOD resources can also be used to provide additional descriptions about the entities. For example, what is Princess Diana's date of birth and death, that she was married to Prince Charles, and so on. Such information can provide an additional context to the users doing a research on a history of Princes Diana, letting their focus on specific events, and not worry about some general characteristics.

The final goal is that this additional data and information gathered together and presented in a nice way will help users to better understand the archive and save their precious time by reducing the amount of articles required to read in order to understand a particular story.

2 Archive Explorer

A news data flow in our system starts with a text mining of the articles one-by-one and creating a database from the newly extracted and contextualized data. The second part provides contextual browsing and querying capabilities of the archive. The architecture can be seen on Fig. 1.

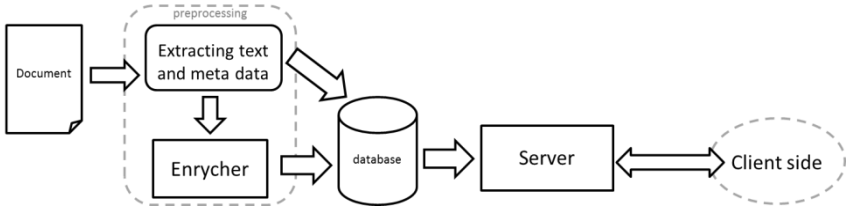

Fig. 1 The architecture of Archive Explorer.

2.1 Extracting data

The system uses a service-oriented framework Enrycher [1] to pre-processes articles. It extracts information with pattern-based and supervised learning knowledge extraction techniques [2] and produces a list of entities, with some of them linked to resources from several LOD dataset: DBpedia, Freebase, New York Times Topics, OpenCyc and Yago. We use the extracted entities and links to LOD to provide infor-

mation about people and organizations involved, information in which cities, countries or other known places events happened and information about other things, like important milestones that are included in viewed articles. Enrycher also provides a taxonomy categorization using DMoz categories, and a set of descriptive keywords.

The output of Enrycher provides a large part of articles' context. These are extended with information extracted from linked LOD resources.

2.2 Browse and Search

Once data is pre-processed and stored it needs to be presented properly. A regular search form is upgraded with a faceted search interface to help the users in browsing around and is appropriate for explorative analysis. If the users know more specifically what they are looking for, they can search across several dimensions and get the search results with all contextual information. One of goals of Archive Explorer is to put a power of the queries and advantages of the visualization together to make context useful and transparent. An application for dynamic re-ranking and visualization of search results Searchpoint [3] is used to let users sort the search results. It uses entities, connections between entities and articles for ranking and ordering.

Fig. 2 Searchpoint visualization of entities. In location window red dot is dragged up to the Brooklyn.

The extracted entities are classified into several types (person, organization, location) and are used to divide the ranking criteria into three different parts of visualization. Every part is presented in its own window and entities are illustrated with spots of different colours. For choosing entity in specific window, users can drag a red dot around with a mouse, which results in a new ranking of articles. The order is changed in a way, so the articles most connected with entities nearest to the red dot are pushed to the top of list of the search results. With this users are narrowing down their search criteria without even knowing this in advance. In Fig. 2 a red dot in a location window is put on the entity New York, in other windows red dots are left at the centre position, which means that articles connected or related to New York are at the top of search results.

2.3 Contextualization and Summarization

A text of news can be short or very long and if we want to show more news articles, it can get really messy. What we want to show is an overview over a collection of articles without losing a user in endless text. Instead of a huge amount of text, we use just information that was extracted and gathered before. If someone is interested in certain topic, he can see which people, locations or other important things are present and how they are connected with each other.

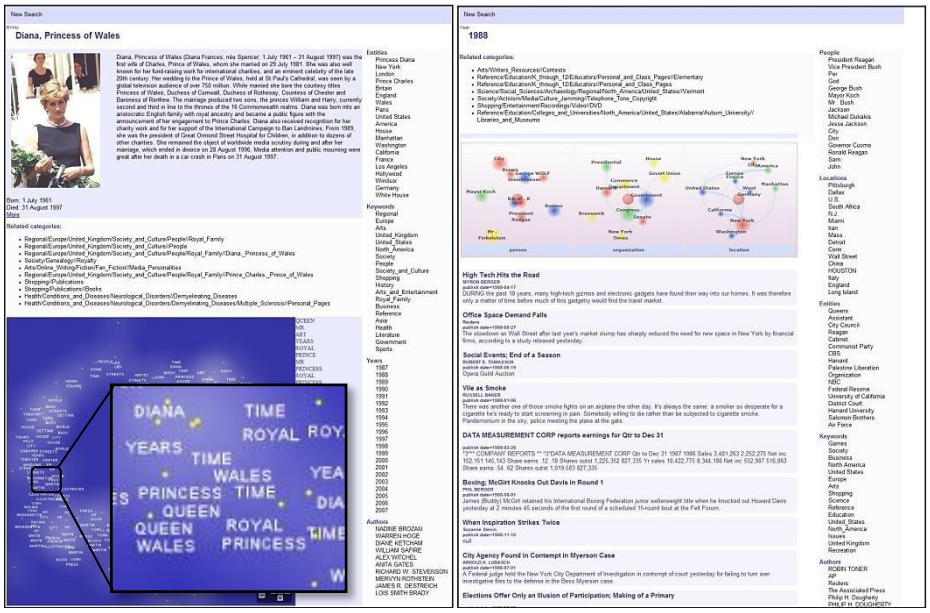

Fig. 3 Connections between entities for search results visualized with graph.

Not just entities, also keywords are handy to visualize connections between contents of articles to understand, discover and summarize the topics in articles. This visualization Document Atlas [4] is used to show the whole picture of search results. With that picture it is also illustrated which articles belong together in the same story or in the same topic. Based on a similarity between documents, they are mapped onto a two-dimensional plan which represents a semantic space of articles and named-entities. Articles having very similar content have coordinates closer to each other than those that are less similar [5].

This visualization is good for bigger groups of articles where it is very essential that we do not put too many information in order to keep things clear and helpful. The same applies for presenting periods of time. From that kind of visualization users can quickly guess main topics that were important during that time. In Fig. 3 on the left picture we can see entities, keywords, authors and other things connected to Princess Diana, including a picture and other information from LOD seen on upper part of the left picture and on the right picture we can see things that were important for the

whole year of 1988.The left part also demonstrates the use of Document Atlas. On a magnified part of it we can see yellow dots presenting articles and keywords presenting topics of nearby articles.

3 Demonstration

Archive Explorer is designed for all types of users. On demonstration, visitors will be able to take different perspectives and try our system either as a historian, a student working on his homework for history class or just a random user curious about events in the past. We will demonstrate usability and point out that a user is not just offered with some options but is getting help to find what he wants and hopefully encouraged to read and search more about related topics. Visitors will be also provided with information about parts of system that cannot be shown on the live demonstration.

4 Conclusions and Future Work

At this point Archive Explorer is a working system providing a framework for including additional visualization and summarization techniques to better show the content hidden in the archives. One particular area for improvement is the time component of news and its visualization. Additionally, search of articles can be improved with narrowing criteria using faceted search and query suggestions.

5 Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under PASCAL2 (IST-NoE-216886), RENDER (ICT-257790-STREP) and PlanetData (ICT-NoE-257641).

References

[1] Enrycher, http://enrycher.ijs.si.
[2] Stajner, T.; Rusu, D.; Dali, L.; Fortuna, B.; Mladenić, D.; Grobelnik, M. A service oriented framework for natural language text enrichment. Informatica 34, 3 (2010).
[3] B. Pajntar, M. Grobelnik. SearchPoint – a New Paradigm of Web Search. 17th International World Wide Web Conference (WWW2008) Developers Track, 2008, http://searchpoint.si.
[4] Document Atlas, http://docatlas.ijs.si.
[5] Fortuna, B.; Mladenić, D.; Grobelnik, M. Visualization of Temporal Semantic Spaces. In: Davies, J. et al (ed.) Semantic Knowledge Management (Springer, 2008).