

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/266891172>

Exploring History Through Newspaper Archives

Conference Paper · April 2015

DOI: 10.1007/978-3-662-46641-4_28

CITATION

1

READS

66

4 authors, including:



Marko Grobelnik

Jožef Stefan Institute

303 PUBLICATIONS 3,944 CITATIONS

[SEE PROFILE](#)



Blaž Fortuna

Jožef Stefan Institute

80 PUBLICATIONS 1,253 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



ad hoc routing protocol. [View project](#)



X5GON: Cross Modal, Cross Cultural, Cross Lingual, Cross Domain, and Cross Site Global OER Network [View project](#)

EXPLORING HISTORY THROUGH NEWSPAPER ARCHIVES

Jasna Škrbec, Marko Grobelnik, Blaž Fortuna, Boštjan Pajntar

Artificial Intelligence Laboratory

Jozef Stefan Institute

Jamova 39, 1000 Ljubljana, Slovenia

Tel: +386 1 4773127; fax: +386 1 4773851

E-mail: jasna.skrbec@ijs.si

ABSTRACT

This paper proposes a pipeline for searching and browsing through newspaper archives. It uses a combination of algorithms for extracting information from text and tools for visualizing different text structures capable of handling large amount of articles that are normally collected in archives. The proposed pipeline is implemented as a web application, illustrative results show appropriateness of the proposed pipeline for searching and browsing news archives.

1 INTRODUCTION

News publishers are collecting old articles into large archives with millions of articles. Even though the articles are typically annotated with additional meta data, the archives are hard to browse or used to discover larger stories. Typically search interfaces do not work well since they are not specialized for archives, and as such do not take advantage of the inherent structure.

This paper presents Archive Explorer, a system for browsing the archives, which combines text mining and visualization techniques. The goal is to go beyond typical search and browse interfaces, which focus on retrieval and visualization of articles, and try to show the articles in context with the rest of the archive. For example, how does an article fit into larger story, which developed over longer time period, and is discussed in many articles, or how is the searched topic represented in the archive with respect to time, place, major events, important people, important keywords etc. The system is designed to get the user's attention and interest for browsing through other related issues.

2 ARCHITECTURE AND BASIC PIPELINE FOR DOCUMENTS IN ARCHIVE EXPLORER

Basic architecture is combined from a pre-processing articles and a live processing of a data from a database as can be seen on Figure 1.

In pre-processing level archives are imported in database from xml files. Xml file contains text and meta data which

varies from archive to archive. Mostly all files have author or authors of article, publish date, title and which part of text is lead paragraph. Usually there are added pages or section where it was published, internal categories, place where it was written, etc. A pure text article and its basic meta data are taken from xml and stored in a database.

Second part of pre-processing is enriching a text with some context. The text from a database is sent to Enrycher application [1]. Enrycher is a service-oriented framework for extraction and representation of document content. Information is extracted from an unstructured document with different knowledge extraction techniques. The first result of Enrycher used by Archive Explorer is extracted list of named entities, such as people, organizations, cities, countries or other places and things that are commonly known. They are extracted from a text with two different techniques, a pattern-based and a supervised learning one [2]. The second result of Enrycher is classification of articles into categories and extraction of keywords. A hierarchical classifier is used for a taxonomy categorization. Relevant categories are defined by a word and a phrase similarity. Hierarchy for categories is taken from DMoz topic ontology [3] and consists of standard centroid vectors. Comparison with document is started by hierarchy's top category and then down through tree. Result is a list of categories with their whole hierarchy path which are ranked by similarity to a document. Only top ten most similar ones are used to classify an article.

All results given by Enrycher are inserted into a database and used to describe a context of an article. With a context it is possible to connect an article with other articles into stories and to place it into timeline.

Articles must be pre-processed and placed in a database before they are used in processing on the server. It is very important to have a rich context around some article if we want to have a good base for connecting and processing the whole archive. Since a typical newspaper archive easily contains millions of articles, the pre-processing time can quickly become a bottle neck of whole system. That is why articles are first saved into database and later they are sent to get a context. So an article can be found even if it has no context around it yet. Faster way to get a context is with

multithreading and speeding up the whole process of enrichment.

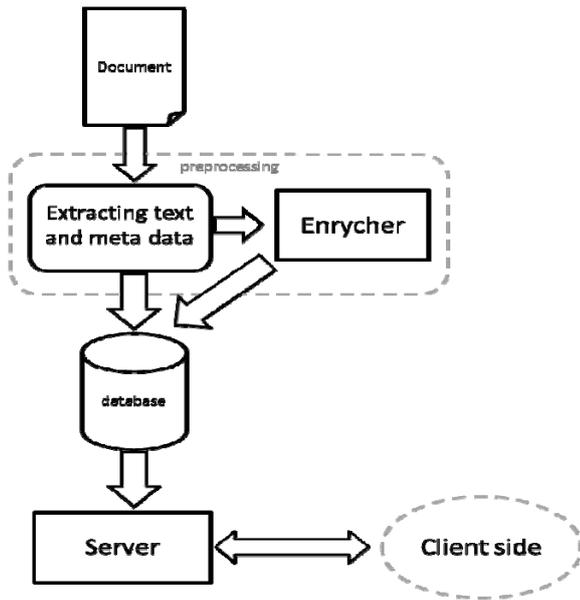


Figure 1: Basic architecture of Archive Explorer.

3 EXPLORING ARCHIVES

If we want to offer archives that are readable and can be explored in interesting and convenient ways, we need to offer more than just an interesting text. Instead of having just a regular search form it can be upgraded with a faceted search interface.

One way for user to start browsing around is to choose entity, keyword, year, author or one of the main categories which are put on first site of the application. The other way, which is most common, is that user type in his search query. In that purpose there are several fields so user can narrow down what he is looking for. Next step is showing user search results with some context around those results. Then a user can choose from given options from a context in which direction the search continues. With offering user options related with his search is not just helping him find what he wants, but also encouraging him to read and search more about this or related topics since there will be only topics he chose before; topics he is interested in.

In an archive with so many articles there are a lot of similar topics. Finding them is not easy especially when user is not sure what or when or who is he looking for. Some nice visualization can manage problems like that. One of goals of Archive Explorer is to put a power of the queries and advantages of the visualization together.

3.1 Browsing search results with Searchpoint

First example of cooperation between a context and a visualisation is Searchpoint [4]. It is an application for ranking and visualizing search results from an ordinary search engine. In our case Searchpoint is using entities,

connections between entities and articles for ranking. Entities are divided into four different groups which are presented in four different parts of visualization.

Every part is presented in its own window where entities are illustrated with a text and spots of different colours. In the middle of every window there is a red dot that can be dragged around with a mouse. If the red dot is moved then order of articles will change. An order is changed in the way so articles most connected with entity or entities nearest to the red dot are on the top of search results. For example if we are interested in a specific person we should drag a red dot in a person window to cover a spot where a name of this person is illustrated. Articles pushed on the top are the ones most relevant for this person.

For testing purposes all examples are from part of New York Times archive. When we typed in search query 'art', Searchpoint returned entities illustrated on Figure 2. In location window we dragged red dot from centre up so the nearest entity/location is Brooklyn, in other windows we leave red dots at the centre position, which means that in search results articles concerning Brooklyn are at the top.

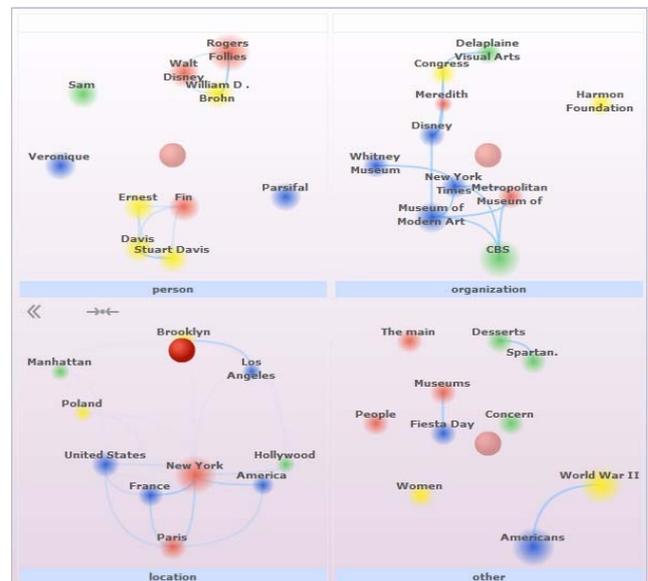


Figure 2: Searchpoint visualization of entities. In location window red dot is dragged up to the Brooklyn.

3.2 Visualization with graphs

All kind of graphs can be also very powerful visualization tool. Sometimes it is enough just to put all set of some keywords on the spot, just to see how many there are. It is very illustrative for articles if there is a picture with all the words used in it where words are bigger if they are used more commonly. Of course it is smart to remove all stop words.

With help of Searchpoint it is obviously which entities are related with which articles, but it is also very interesting how and which entities are related between themselves. So if someone is interested in some topic, he can see which person is connected with other person or with which city or

which city has the most connections in this topic, etc. JavaScript library called Arbor.js [5] is used for visualization of that kind of a graph. For now graphs are used to show connection between entities found in search results and to illustrate one entity's connections. In Figure 3 we can see example for first case. Data needed to realize graphs is provided by sql queries and server live processing.

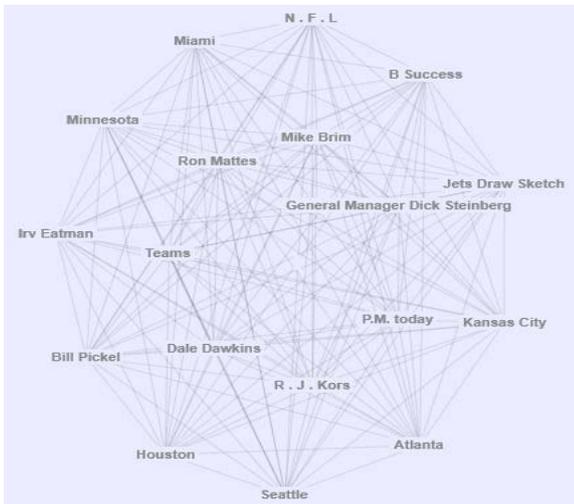


Figure 3: Connections between entities for one article visualized with graph.

3.3 Visualization of article's content

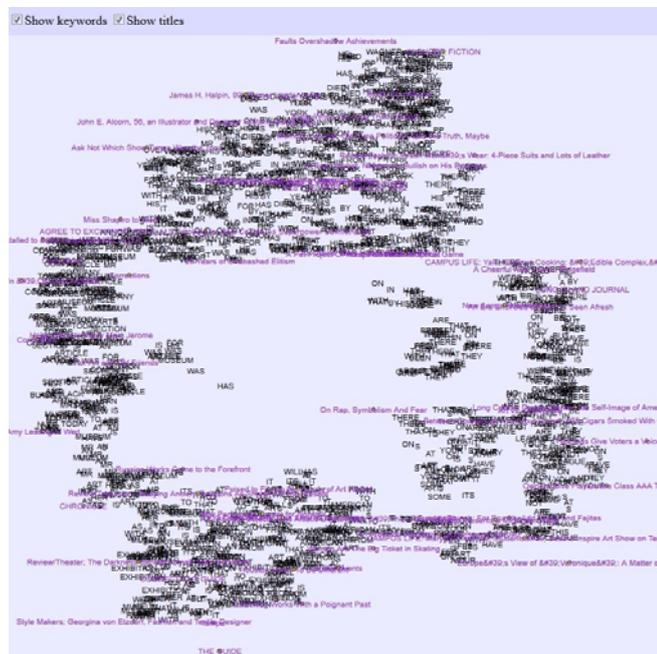


Figure 4: Document Atlas visualization of articles.

Visualizations described in last two sections are meant for showing how meta data is connected with either other meta data either with articles. Visualization in this section is intended for connections between contents of articles. It

helps to understand, discover and summarize the topics in articles. This visualization is called Document Atlas [6] and example of use in our system is shown in Figure 4.

Documents pushed into an application are mapped onto a two-dimensional plane based on similarity between them. If documents are very similar by content, then their coordinates are closer than those of documents that are less similar. A set of different methods is used for a calculation of the semantic space which includes documents and named-entities from the text corpus [7]. In Archive Explorer Document Atlas is used to show the whole picture of search results. With that picture it is illustrated which articles belong together in same stories or in same topics. On picture there are article's titles and their keywords. From set of keywords we can guess topics on different spots.

3.4 Timeline

People, places, organizations, major events in history and others such things are covered with entities, keywords and other context, but time perspective is missing. Since we are working with archives a time component is quite important. For example, if someone is searching for a specific city in an archive, he would probably want to have some overview by its history over the years. One way to visualized search results is also by years or by months. Graph visualized with help of Raphael library [8] shows number of found articles by years. A set of articles in the chosen year is best shown with Document Atlas. The same kind of graph is also used for showing numbers of articles per month for one year (Figure 5). This is useful for spotting important events in a specific year.



Figure 5: Number of articles per month for one year.

4 EXAMPLE FROM USER POINT OF VIEW

Let's say we have a student who has to do a research for his history class. In research he has to decide how different events have influenced its history. He chooses a city where he was born and focuses on the time when he was born. So he decides to find some bigger newspaper which existed back than and has online archive. Since he is interested in specific date and location he type in his query, for example in location text field he type New York and in date field he type 1992-02-02.

When he gets results (Figure 6) he can immediately see from Searchpoint which people and organizations were the most important at that time in the city and he can see connections with other locations. He is interested if the

