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 shows 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.

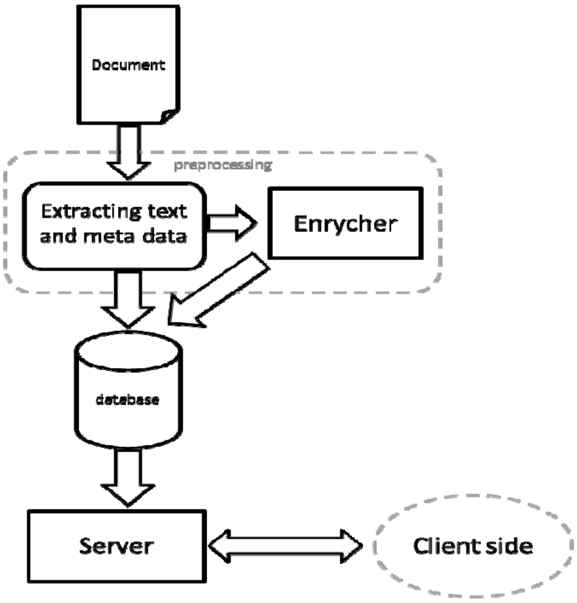


Figure 1: *Basic architecture of Archive Explorer.*

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.

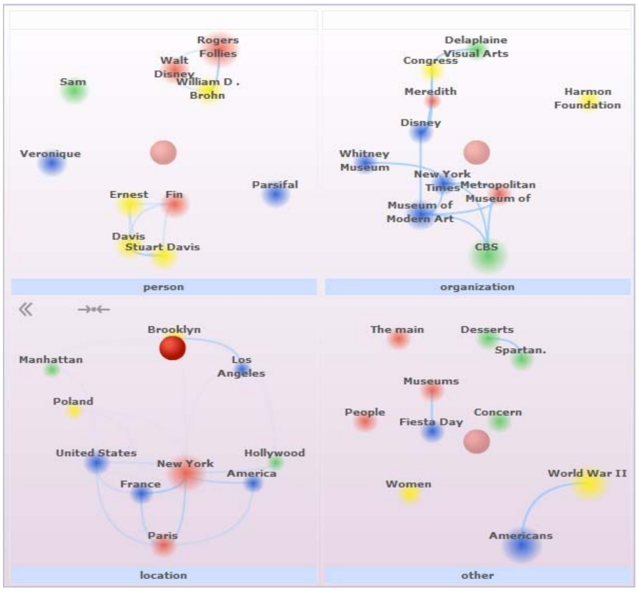


Figure 2: *Searchpoint visualization of entities. In location window red dot is dragged up to the Brooklyn.*

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.

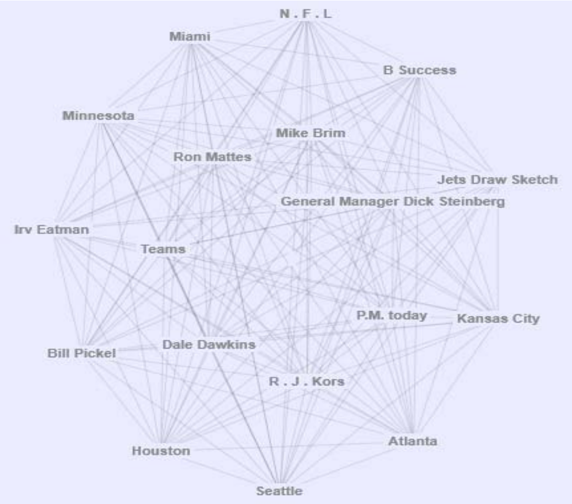


Figure 3: *Connections between entities for one article visualized with graph.*

3.3 Visualization of article's content

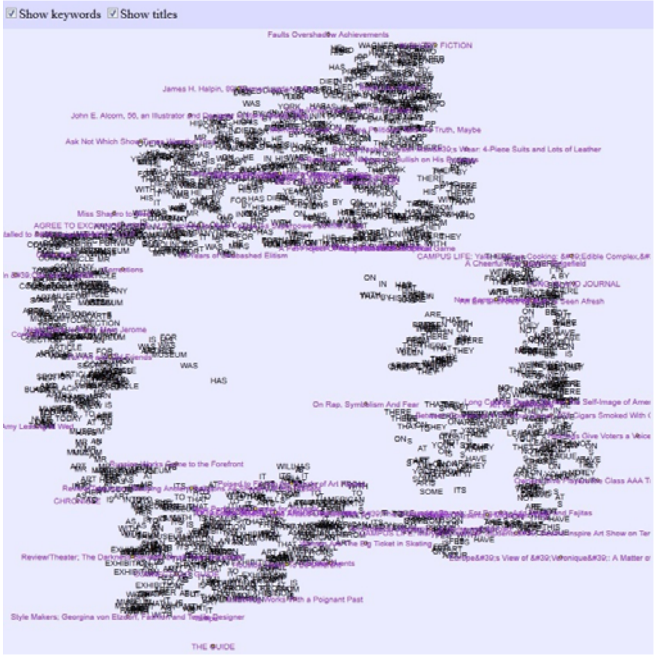


Figure 4: *Document Atlas visualization of articles.*

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 Raphel 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.

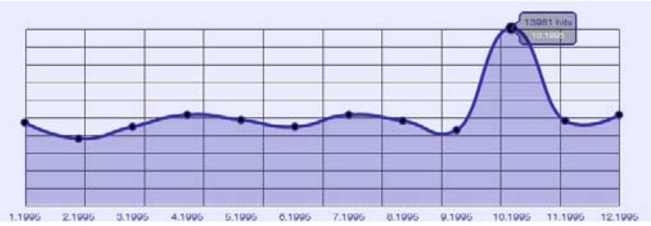
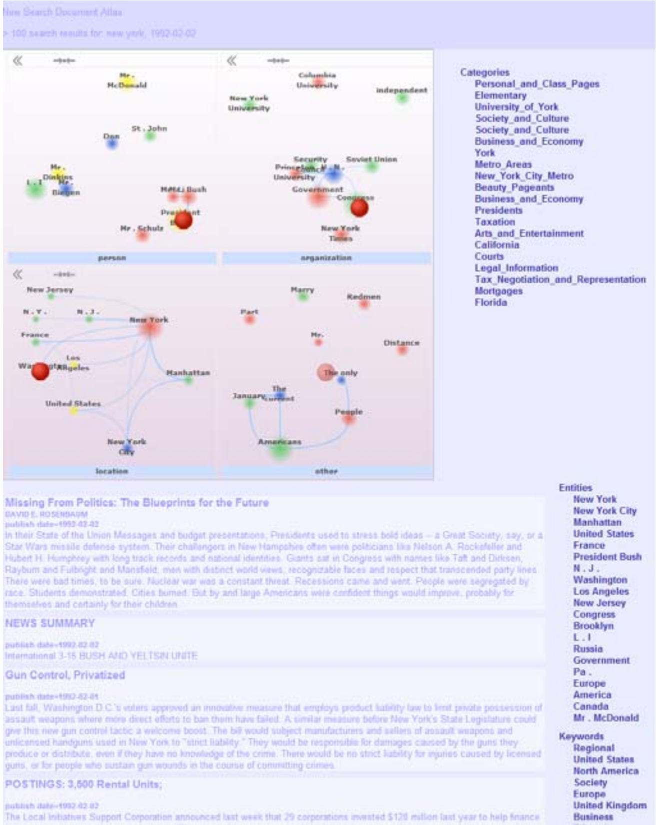


Figure 5: *Number of articles per month for one year.*

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 events were new or old, so he browses forward with the help of sidebars to find important categories, entities, keyword and authors related with the city in the specified time period.



New Search Document Atlas

> 100 search results for: new york, 1992-02-02

Categories
- Personal_and_Class_Pages
- Elementary
- University_of_York
- Society_and_Culture
- Society_and_Culture
- Business_and_Economy
- York
- Metro_Areas
- New_York_City_Metro
- Beauty_Pageants
- Business_and_Economy
- Presidents
- Taxation
- Arts_and_Entertainment
- California
- Courts
- Legal_Information
- Tax_Negotiation_and_Representation
- Mortgages
- Florida

Missing From Politics: The Blueprints for the Future

DAVID E. ROSENBAUM

publish date=1992-02-02

In their State of the Union Messages and budget presentations, Presidents used to stress bold ideas -- a Great Society, say, or a Star Wars missile defense system. Their challengers in New Hampshire often were politicians like Nelson A. Rockefeller and Hubert H. Humphrey with long track records and national identities. Giants sat in Congress with names like Taft and Dirksen, Rayburn and Fulbright and Mansfield, men with distinct world views, recognizable faces and respect that transcended party lines. There were bad times, to be sure. Nuclear war was a constant threat. Recessions came and went. People were segregated by race. Students demonstrated. Cities burned. But by and large Americans were confident things would improve, probably for themselves and certainly for their children.

NEWS SUMMARY

publish date=1992-02-02

International 3-15 BUSH AND YELTSIN UNITE

Gun Control, Privatized

publish date=1992-02-01

Last fall, Washington D.C.'s voters approved an innovative measure that employs product liability law to limit private possession of assault weapons where more direct efforts to ban them have failed. A similar measure before New York's State Legislature could give this new gun control tactic a welcome boost. The bill would subject manufacturers and sellers of assault weapons and unlicensed handguns used in New York to "strict liability." They would be responsible for damages caused by the guns they produce or distribute, even if they have no knowledge of the crime. There would be no strict liability for injuries caused by licensed guns, or for people who sustain gun wounds in the course of committing crimes.

POSTINGS: 3,500 Rental Units;

publish date=1992-02-02

The Local Initiatives Support Corporation announced last week that 29 corporations invested $128 million last year to help finance

Entities
- New York
- New York City
- Manhattan
- United States
- France
- President Bush
- N . J .
- Washington
- Los Angeles
- New Jersey
- Congress
- Brooklyn
- L . I
- Russia
- Government
- Pa .
- Europe
- America
- Canada
- Mr . McDonald

Keywords
- Regional
- United States
- North America
- Society
- Europe
- United Kingdom
- Business

Figure 6: *Page with search results and Searchpoint.*

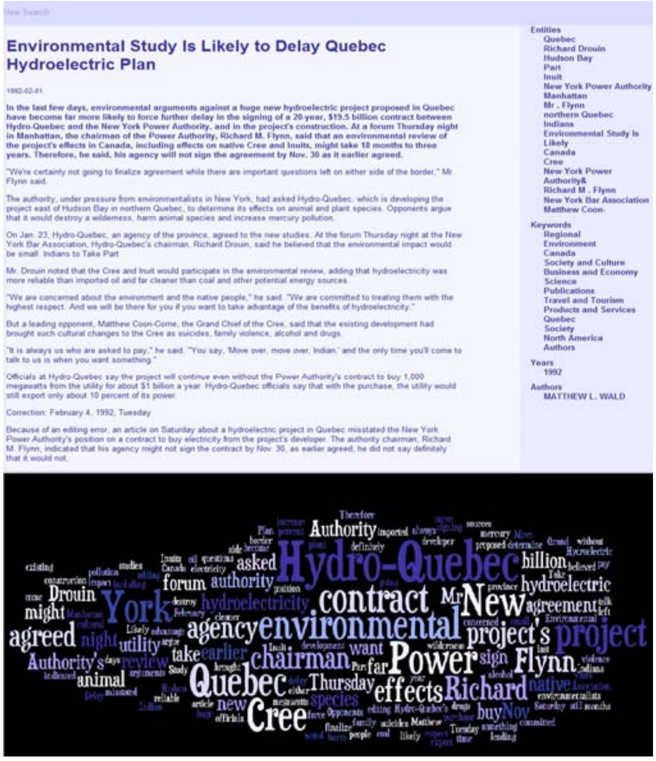
By browsing authors' pages he gets image of what kind of publisher is specific author. So if he found a lot of articles from author who was writing mostly about art, he knows there were a lot of movement in this area by that time.

Because he doesn't want to read everything he looks over Document Atlas to see which were the major topics at the time. There he discovers that in that time they were planning to build hydroelectric plan on border with Canada. Now he can go further browsing about that. The article about hydroelectric plan is on Figure 7.

5 CONCLUSIONS AND FUTURE WORK

At this point on one side Archive Explorer is a working system but on the other side it is just a basic scheme of possibilities that can be implemented in a system like that. It already shows the advantages of merging text mining and visualization.

A lot of ideas are still to be implemented. For example, some existing visualizations could be used in more areas of the system. Search of articles can be improved with narrowing criteria and query suggestions. We will also try to automatically connect articles into stories. The time component has also has a lot of space for improvement.



New Search

Environmental Study Is Likely to Delay Quebec Hydroelectric Plan

1992-02-01

In the last few days, environmental arguments against a huge new hydroelectric project proposed in Quebec have become far more likely to force further delay in the signing of a 20-year, $19.5 billion contract between Hydro-Quebec and the New York Power Authority, and in the project's construction. At a forum Thursday night in Manhattan, the chairman of the Power Authority, Richard M. Flynn, said that an environmental review of the project's effects in Canada, including effects on native Cree and Inuits, might take 18 months to three years. Therefore, he said, his agency will not sign the agreement by Nov. 30 as it earlier agreed.

"We're certainly not going to finalize agreement while there are important questions left on either side of the border," Mr. Flynn said.

The authority, under pressure from environmentalists in New York, had asked Hydro-Quebec, which is developing the project east of Hudson Bay in northern Quebec, to determine its effects on animal and plant species. Opponents argue that it would destroy a wilderness, harm animal species and increase mercury pollution.

On Jan. 23, Hydro-Quebec, an agency of the province, agreed to the new studies. At the forum Thursday night at the New York Bar Association, Hydro-Quebec's chairman, Richard Drouin, said he believed that the environmental impact would be small. Indians to Take Part

Mr. Drouin noted that the Cree and Inuit would participate in the environmental review, adding that hydroelectricity was more reliable than imported oil and far cleaner than coal and other potential energy sources.

"We are concerned about the environment and the native people," he said. "We are committed to treating them with the highest respect. And we will be there for you if you want to take advantage of the benefits of hydroelectricity."

But a leading opponent, Matthew Coon-Come, the Grand Chief of the Cree, said that the existing development had brought such cultural changes to the Cree as suicides, family violence, alcohol and drugs.

"It is always us who are asked to pay," he said. "You say, 'Move over, move over, Indian,' and the only time you'll come to talk to us is when you want something."

Officials at Hydro-Quebec say the project will continue even without the Power Authority's contract to buy 1,000 megawatts from the utility for about $1 billion a year. Hydro-Quebec officials say that with the purchase, the utility would still export only about 10 percent of its power.

Correction: February 4, 1992, Tuesday

Because of an editing error, an article on Saturday about a hydroelectric project in Quebec misstated the New York Power Authority's position on a contract to buy electricity from the project's developer. The authority chairman, Richard M. Flynn, indicated that his agency might not sign the contract by Nov. 30, as earlier agreed; he did not say definitely that it would not.

Entities
- Quebec
- Richard Drouin
- Hudson Bay
- Part
- Inuit
- New York Power Authority
- Manhattan
- Mr . Flynn
- northern Quebec
- Indians
- Environmental Study Is Likely
- Canada
- Cree
- New York Power Authority&
- Richard M . Flynn
- New York Bar Association
- Matthew Coon-

Keywords
- Regional
- Environment
- Canada
- Society and Culture
- Business and Economy
- Science
- Publications
- Travel and Tourism
- Products and Services
- Quebec
- Society
- North America
- Authors

Years
- 1992

Authors
- MATTHEW L. WALD

Figure 7: *Article's page.*

6 ACKNOWLEDGMENTS

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

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] DMoz, http://www.dmoz.org

[4] Searchpoint, http://searchpoint.si.

[5] Arbor.js, http://arborjs.org.

[6] Document Atlas, http://docatlas.ijs.si.

[7] Fortuna, B.; Mladenić, D.; Grobelnik, M. Visualization of Temporal Semantic Spaces. In: Davies, J. et al (ed.) Semantic Knowledge Management (Springer, 2008).

[8] Raphael, http://raphaeljs.com/