Gaining insights into the Global Financial Crisis using Saffron

Georgeta Bordea **Kartik Asooja** **Paul Buitelaar**
INSIGHT Centre for Data Analytics
National University of Ireland, Galway
name.surname@insight-centre.org

Leona O'Brien
GRCTC, University College Cork, Ireland

Abstract

We participated in the NLP Unshared Task in PoliInformatics 2014 with Saffron, a system that provides insight into the Financial Crisis by analysing main topics of discussion and experts associated with these topics. Saffron applies term extraction, topical hierarchy construction, expert finding and expert profiling to provide a set of tools that allow users to investigate what was the financial crisis and who was the financial crisis.

1 Introduction

The financial crisis of 2007-2008 was a complex series of events that led to world-wide recession, with implications that are still unravelling in many sectors of the real economy. The response of the U.S. government to this crisis was documented in a large number of reports, hearings, bills, and transcripts that provide a rich resource of study for researchers in political science and communications. The sheer size and complexity of this corpus is a challenge, requiring natural language processing techniques to extract and locate relevant information about lawmaking and regulatory processes and the people involved in these processes.

In this paper, we propose using Saffron [1] as a tool for exploratory search of topics, documents and people related to the financial crisis. The NLP Unshared Task in PoliInformatics 2014 provided participants with a corpus on the financial crisis and defined the following broad research questions:

1. Who was the financial crisis?

2. What was the financial crisis?

We cast the first research problem as an expert finding task, similar to previous work done for locating knowledgeable people inside a large enterprise (Bordea et al., 2012). Expert finding is the task of ranking people that have knowledge about a given topic by performing a full text search for experts instead of documents. The information retrieval community encouraged research on expert finding by organising several shared tasks (Craswell et al., 2006; Soboroff et al., 2007; Bailey et al., 2007) as part of the Text REtrieval Conference (TREC)[2]. In the context of the financial crisis, this allows users to identify individuals that were directly involved in defining the response of the U.S. government to the financial crisis by searching for a topic of interest.

Additionally, we consider the expert profiling task, which addresses the problem of constructing concise descriptions about a person by identifying associated expertise topics. This functionality provides users with more context about proposals, arguments, and policies on which a person is an expert on. Expert profiling is an important component of an expert finding system, but previously it was assumed that a list of relevant knowledge areas is available (Balog and de Rijke, 2007). We address the problem of discovery and identification of expertise topics as part of the second research question from this task.

Saffron provides tools for investigating what was the financial crisis through term extraction and automatic construction of topical hierarchies. Term extraction is a central component in Saffron, with a focus on extracting terms of an intermediate level of specificity, which are useful for summarisation and classification (Bordea et al., 2013). Although informative, flat lists of terms can only provide limited insight in a domain, while topical hierarchies organise extracted terms in an intuitive structure, summarising content and en-

[1] Demo: http://saffron.deri.ie/finance

[2] http://trec.nist.gov/

abling exploratory access to information. Saffron automatically constructs a topical hierarchy from a domain-specific corpus, organising terms from general terms to more specific terms.

This paper is organised as follows. First, we provide an overview of the overall Saffron architecture in Section 2, then we discuss some of the key insights about the financial crisis that can be gathered using Saffron in Section 3. We conclude this paper with a discussion in Section 4.

2 The Saffron system

In this section, we provide an overview of the Saffron infrastructure stack, depicted in Figure 1. The first stage is related to the acquisition of a domain corpus, followed by a preprocessing stage. The main layer of the architecture is the Expertise Mining layer, that extracts expertise information which is then passed to the Storage and Index layer, before it is visualised on the Frontend.

Data acquisition and data preprocessing

Saffron was previously applied in various application scenarios, including finding experts in an enterprise and locating collaborators during academic events. Therefore, various input formats were considered when gathering suitable datasets about individuals and associated documents. Metadata about authors is most often represented in XML, but there is an increasing number of datasources available in RDF. RESTful Web Services are another way to provide access to expertise datasets.

In the case of academic events such as conferences and workshops, it is often the case that information about publications and authors is not readily available, and has to be collected from dedicated HTML websites through web scraping. In the enterprise environment, most information about documents and organisation members is not public, and has to be accessed from content management tools, such as SharePoint. Saffron uses a popular open source relational database, MySQL, as backend, and Lucene, an information retrieval library, for indexing full-content documents.

Expertise Mining

This layer addresses the core tasks of Expertise Mining, including expertise topic extraction, topical hierarchy construction, expert profiling, and expert finding. Candidate terms are identified using a NLP pipeline based on the GATE natural language processing framework (Cunningham et al., 2002) and the ANNIE information extraction system, included in the standard GATE distribution.

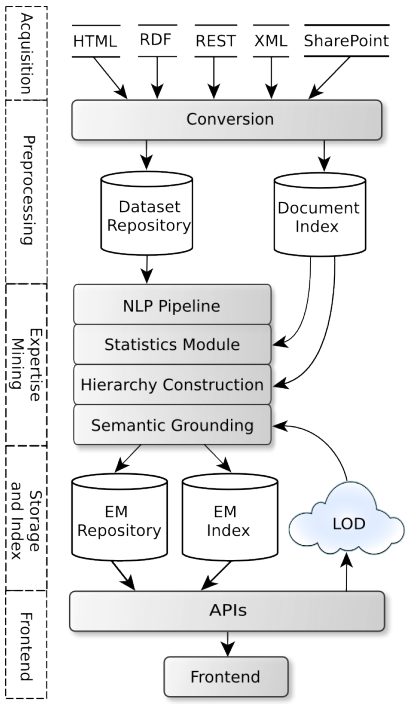

Figure 1: Overview of the Saffron infrastructure stack

The Statistics Module is responsible for ranking and filtering candidate terms to identify expertise topics. Word occurrences, as well as relevance measures for expert finding and expert profiling, are computed using the Lucene index. The relations between expertise topics are identified by the Hierarchy Construction component, using a graph-based algorithm for constructing topical hierarchies described in (Hooper et al., 2013). Again, Saffron relies on the Lucene index to measure co-occurrences between two expertise topics, making use of the span search functionality available in Lucene.

The final core Expertise Mining component, Semantic Grounding, is responsible for identifying DBpedia URIs and descriptions of expertise topics. In this way Saffron provides an entry point in the Linked Open Data cloud, as well as descriptions for expertise topics that can be directly used by the end user.

Expertise storage and index

The Storage and Index layer prepares the data for high performance access. This is done either directly through APIs or through a SPARQL end-

point. The EM Repository is a MySQL based solution for storing data. The Expertise Mining results are also indexed by the EM Index component, using Solr, a higly scalable enterprise search engine.

3 Using Saffron for analysing the Global Financial Crisis

Saffron is deployed as a web application, displaying on the start page a short list of top ranked terms, and a topical hierarchy which allows users to discover more specific topics, as can be seen in Figure 2. Each node from the topical hierarchy can be used to browse information about a specific topic. A standard search interface is also provided, allowing users to look for a particular person, topic or document. In the rest of the section we provide a brief description of the financial crisis corpus used in our analysis and then we discuss some of the insights that can be gathered from this corpus using Saffron.

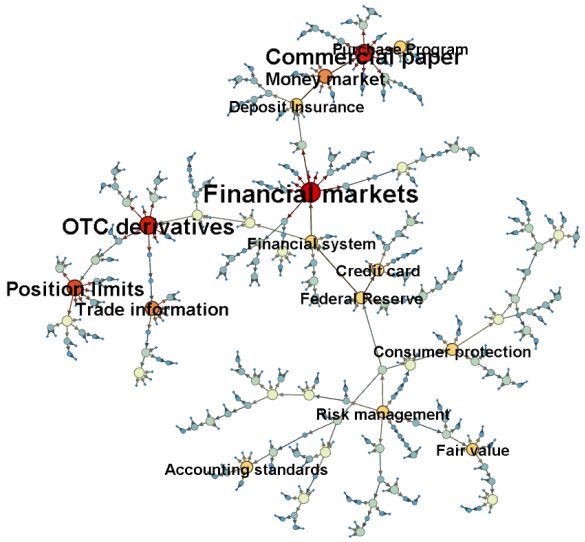

Figure 2: Topical hierarchy used for browsing the financial crisis documents

3.1 Financial crisis corpus

The task organisers provided participants with a corpus of reports, hearings, bills, and other transcripts related to the financial crisis. Data sources included the Federal Open Market Committee (FOMC), the Federal Crisis Inquiry Commission (FCIC), Congressional reports, Congressional bills, and Congressional hearings. In our analysis, it was only the official collection of the task that was considered.

Metadata about publishing year and people involved in the development of the document was extracted directly from the text of the document, where available. We were able to extract metadata for the following types of the documents: FOMC, FCIC, Congressional Hearings, and Congressional Reports. It was assumed that each person had the same contribution in producing the content of a document. A more sophisticated approach where each utterance is identified and associated with the corresponding person can be envisioned but we leave this direction for future work.

For the Congressional Hearings documents and the FOMC documents, we defined regular expressions to extract the year and the full names of the witnesses and committee members from each document. We could not obtain full names of authors in the FOMC documents, because these documents do not contain this information. In the case of the FCIC documents and of the Congressional Reports, the metadata was collected manually from the .pdf files, as there were just two documents in each case.

3.2 Who was the financial crisis?

Saffron allows users to identify people that were involved in the debate around specific topics as can be seen in Figure 3. This example illustrates the expert finding scenario, where Saffron can be used to identify experts on a given topic. Saffron identifies the politicians Chris Dodd and Mel Watt, representing the Democratic Party, and Spencer Bachus, representing the Republican Party, to be the most involved people in the *Credit Unions* topic. The results of the expert profiling task are presented in Figure 4. We choose the profile of the economist Paul Volcker as an example. The main topic associated with this expert is ”proprietary trading”. These proposals were actually named ”The Volcker Rule”, because he has been the main speaker on the necessity to limit/ban this type of trading. The Volcker Rule regulations were approved in December 2013 by the five regulatory agencies involved, coming into force on 1st April 2014.

3.3 What was the financial crisis?

The topical hierarchy provided on the start page of Saffron allows users to understand the main topics of discussion and the relations between them. The most broad topic from this hierarchy, shown in Figure 2, is the *Risk management* topic, which

Figure 3: Top experts for the Credit Unions topic in the context of the financial crisis

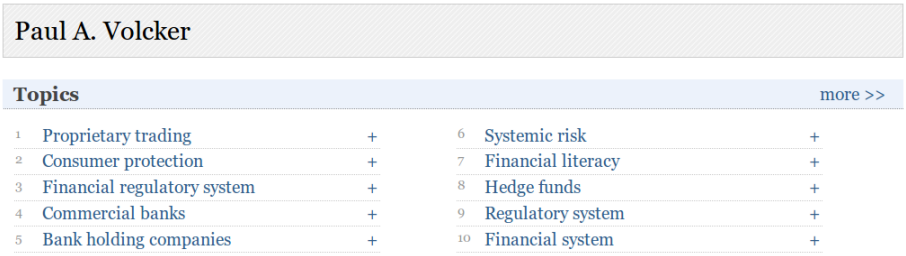

Figure 4: Topical profile of Paul A. Volcker in the context of the financial crisis

is the root of the hierarchy and the most central topic for the U.S. response to the financial crisis.

The hierarchy can be used to discover more fine-grained topics such as *Consumer protection*, which can be seen on the middle-right part of the hierarchy. The U.S. government pushed for stronger Consumer Protection, as the crisis highlighted several shortcomings in this area. The interface can be used to zoom in on a specific area, as can be seen in Figure 5. This visualisation allows the user to discover other related topics such as the creation of a new Consumer Protection Agency, that was meant to reinforce a stronger approach to Consumer Protection.

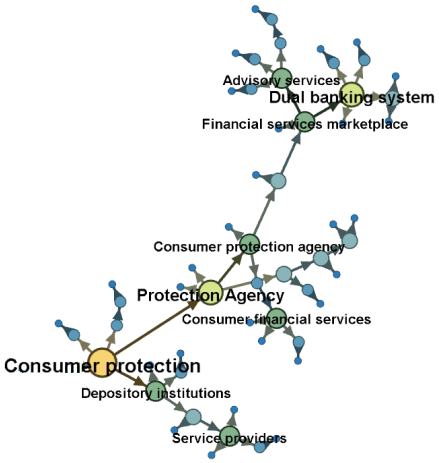

Figure 5: The Consumer Protection subtree from the Financial Crisis topical hierarchy

4 Discussion and Conclusions

In this paper we proposed using Saffron as a tool for gaining insights into the financial crisis. We showed that the question about who was the financial crisis can be framed in terms of expert finding and expert profiling. Also, we presented an exploratory interface of financial crisis topics that makes use of an automatically-constructed topical hierarchy. This interface can be used to get a high-level overview of what was the financial crisis. Although Saffron does not explicitly distinguish between causes, proposals, and policies, each expertise topic is presented together with snippets from the documents where it was mentioned. This allows a user to infer more specific information about the type of a topic. Another limitation of this proof-of-concept is that all the people involved in creating a document were assumed to have the same contribution in producing the whole content. This is an oversimplification, and a solution that associates each utterance separately is likely to improve the results.

Encyclopedic information was only considered for topic descriptions, but background knowledge can be useful for people as well, because the majority are well-known politicians or economists with their own wikipedia page. Other relevant information about a person such as political party could be useful in this context. Future work will make use of the topical hierarchy to visualise expertise profiles in a wider context.

Acknowledgements

This work has been funded in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT) and by Enterprise Ireland (EI) as part of the project Financial Services Governance, Risk and Compliance Technology Centre (GRCTC), University College Cork, Ireland.

References

Peter Bailey, Nick Craswell, Arjen P de Vries, and Ian Soboroff. 2007. Overview of the trec 2007 enterprise track draft. In *TREC 2007 Working notes*.

Krisztian Balog and Maarten de Rijke. 2007. Determining expert profiles (with an application to expert finding). In *proc. of the International Joint Conferences on Artificial Intelligence (IJCAI 2007)*.

Georgeta Bordea, Sabrina Kirrane, Paul Buitelaar, and Bianca O Pereira. 2012. Expertise mining for enterprise content management. In *LREC*, pages 3495–3498.

Georgeta Bordea, Paul Buitelaar, and Tamara Polajnar. 2013. Domain-independent term extraction through domain modelling. In *the 10th International Conference on Terminology and Artificial Intelligence (TIA 2013), Paris, France*. 10th International Conference on Terminology and Artificial Intelligence.

Nick Craswell, Arjen P de Vries, and Ian Soboroff. 2006. Overview of the trec-2005 enterprise track. In *The fourteenth Text REtrieval Conference Proceedings (TREC 2005)*.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Clare J Hooper, Georgeta Bordea, and Paul Buitelaar. 2013. Web science and the two (hundred) cultures: Representation of disciplines publishing in web science. In *Proceedings of Web Science 2013*.

Ian Soboroff, Arjen P de Vries, and Nick Craswell. 2007. Overview of the trec 2006 enterprise track. In *The fifteenth Text REtrieval Conference Proceedings (TREC 2006)*.