Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food

Fergal Monaghan, Georgeta Bordea, Krystian Samp, and Paul Buitelaar

Unit for Natural Language Processing
Digital Enterprise Research Institute, National University of Ireland, Galway
{fergal.monaghan,georgeta.bordea,krystian.samp,paul.buitelaar}@deri.org
http://nlp.deri.ie

Abstract. Saffron is an application that provides users valuable insight into a research community or organisation. It makes use of several heterogeneous information sources that are under diverse ownership and control: it combines structured data from various sources on the Web with information extracted from unstructured documents using Natural Language Processing techniques to show the user a personalised view of the most important expertise topics, researchers and publications. Saffron also applies semantic technology in a novel way that goes beyond pure information retrieval: the system recommends mutual contacts (both professional and social) to the user, who would be able to broker a meaningful "shortcut" introduction to an expert. An explicit design process has resulted in an attractive and functional Web interface which provides users with an experience that goes beyond a research prototype. Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies and validate the results obtained.

Keywords: Interesting to run topic extraction over the paper for these?[1]

1 Introduction

Saffron[2] is an end-user application that provides Web users valuable insight into a research community or organisation: in the case of this demo, this is the Semantic Web research community. It supports the user to get up to speed with an expertise topic area, understand the community or constituent organisations and discover experts of relevance. Saffron always shows the user the most important expertise topics, researchers and publications. The user can navigate through linked descriptions of these, and Saffron ensures that they always see the most relevant expertise topics, researchers and publications corresponding to their selection. The user can also combine selections and find answers to specific questions, or can use search if they have an idea of what they are looking for.

[1] Semantic poop: Saffron ate its own dog food to provide these keywords by extracting expertise topics from this paper.

[2] http://saffron.deri.ie/

Importantly, traditional expert finding systems only output the experts deemed relevant and then leave the user to begin cold calling potentially busy strangers if they have an expertise need. Saffron goes beyond this role to actual expert contacting, by recommending mutual contacts (both professional and social) shared by the user and experts who could act as brokers to make an introduction. In this regard, Saffron provides an automatic, social semantic Yellow Pages directory service. This supports the user to actually connect with the right people in a meaningful way and to act on their expertise need. Some example users and use cases of Saffron would be:

- A new Ph.D. student trying to find their research direction or supervisor
- An entrepreneur looking for a domain expert to form a start-up
- A researcher looking for an expert in another area to form a collaboration
- A new employee who wants to find out more about their organisation

2 Expertise Topic Extraction

Saffron's functionality is different from and goes beyond pure information retrieval. The expertise topics are identified in the text based on their context using a list of manually identified skill types (i.e. high level concepts of a domain) that introduce an expertise topic. The initial candidates are the noun phrases introduced by skill types and then a combination of statistical measures is used to find the expertise topics. We use an adaptation of the standard information retrieval measure TF-IDF to build expertise profiles for researchers.

Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies and validate the results obtained. We evaluate our results both at the expertise topic extraction level, by comparison with keyword extraction baselines and at the expert profile construction level by introducing a benchmark dataset. By participating in the SemEval 2010 competition, in the task "Automatic Keyphrase Extraction from Scientific Articles" [4] we assign the keyphraseness of our expertise topics. The performance of our system was consistently above the baselines [2] and our system was ranked 8th out of 19 participants. The evaluation dataset for profile construction is gathered manually from the data about workshop committee members, assuming that their selection is based on human judgment [3].

Saffron uses contextual information for ranking of results according to context. The document context is used to identify expertise topics by considering as candidates only the noun phrases that are either introduced by a skill type or that contain a skill type as a head noun. We also analyze the context of an expertise topic on a Web scale, by applying a filter based on occurrences on webpages. We measure the relation strength between a researcher and the expertise topics from its expertise profile by using the Sindice Semantic Web search engine. All the expertise topics presented in the interface are ranked, first

at the extraction step based on information from the documents and then based on their association with researchers.

3 Information Sources

Saffron makes use of several information sources that are under diverse ownership and control. First, the Semantic Web Dog Food (SWDF) Corpus[3] provides information on papers that were presented, people who attended, and other things that have to do with the main conferences and workshops in the area of Semantic Web research. Implicit relationships between concepts can also be inferred from SWDF e.g. co-authorship relationships between researchers.

Second, information extracted from SWDF publications source PDF files using Natural Language Processing (NLP) techniques provides expertise topics and weighted relationships between expertise topics, publications and their authoring researchers. In our system, an expertise topic is the name of a scientific area or technology (e.g. "social network", "information retrieval", "image processing", "statistical machine learning"). We also extract expertise evidence from the Semantic Web by building a query containing the quoted full person name and the quoted expertise topic. This query is sent to the Sindice search engine[4] [7] and the returned number of hits is considered as a measure of expertise of a researcher for an expertise topic.

Third, DBPedia is used to provide URIs and descriptions of extracted expertise topics. Fourth, extended information about people is crawled from seed URLs in SWDF in the following manner:

1. All URLs given as the seeAlso link from people were collected.
2. Any triple data available at the collected URLs were crawled. This was carried out twice, so that essentially two levels of depth from SWDF were crawled through. These include information from OntoWiki[5] as well individual FOAF profiles and provide further details on researchers e.g. profile pictures and social network connections.
3. The merge of SWDF and potentially inconsistent crawled data is consolidated quickly using CanonConsolidator [5, ch. 5].

The information sources used are syntactically, structurally and semantically heterogenous. On hand, the publication documents are essentially unstructured syntactical strings and hold no explicit semantics. On the other hand, NLP adds some semantics by extracting expertise topics. Structure is also added by the assignment of weighted relationships between extracted expertise topics, the documents they appear in and those documents' authors. Additionally, while the triples from SWDF, DBPedia and those crawled from

[3] `http://data.semanticweb.org/`
[4] `http://www.sindice.com/`
[5] `http://ontowiki.net/Projects/OntoWiki`

OntoWiki and FOAF profiles are structured similarly, they hold heterogeneous semantics in the differing schemas they employ. They also may be formatted with differing syntax (XML, N-Triples, Turtle), although contemporary RDF parsers make this a relatively trivial distinction for Saffron's purposes.

The information sources used contain substantial quantities of real world data. We perform expertise topic extraction on papers from Semantic Web conferences from 2006-2010. While each paper in SWDF has an identifying URI, not all have a URL link to a corresponding PDF file with the actual paper content. So extraction is performed only on that subset of 747 papers that have such links to PDF files. Table 0(a) gives further numbers on the extraction process, such as the total no. of tokens extracted, the total no. of unique researchers who authored the processed papers, and the total no. of expertise topics identified. These numbers give an average of 320 expertise topics per document and 142 topics per researcher. Furthermore, Table 0(b) gives numbers on the triple data crawled from the Linked Open Data (LOD) Web.

(a) Corpus numbers

tokens	papers	people	topics
5,285,870	747	2,191	45,715

(b) Linked Data numbers (28/9/2010)

	triples	papers	people	knows
swdf	91,241	1,589	3,812	0
crawl1	105,325	1,604	4,664	858
crawl2	141,753	1,854	6,941	3,296
consol.	140,649	1,854	5,513	2,660

Table 1. Dataset numbers

4 The Role of Meaning

The meaning of data plays a central role in Saffron. Meaning is represented using Semantic Web technologies. The meaning of the SWDF and crawled data is represented using RDF, RDFS and OWL ontologies. In particular, Inverse Functional Properties (IFPs) represented in OWL ontologies are used to consolidate the crawled data about researchers. Additionally, SPARQL is used to query the data in a powerful, expressive and meaningful way.

Furthermore, Saffron attempts to assign each extracted expertise topic to a concept URI from the LOD Web. In the current prototype we search only for URIs from the DBPedia domain. For each expertise topic we build a query containing the quoted expertise topics and we analyze the first 10 results retrieved by Sindice. We associate the expertise topic with a URI by performing a string based comparison with the title of the webpage and the URI link itself. In this manner we associated 1,823 extracted expertise topics with DBpedia concepts.

Finally, while extracted expertise topics exist within Saffron and are output to the user via the UI, future work aims to explicitly encode all extracted expertise topics and their relationships to papers and researchers as RDF, effectively extending the SWDF dataset.

Saffron manipulates and processes data in interesting ways to derive useful information, and has novelty in applying semantic technology to a domain and task that has not been considered before. It automatically extracts expertise topics from papers, assigns expertise to researchers and connects to existing structured data on the LOD Web. While submissions to previous Semantic Web Challenges, such as RKBExplorer[6] and SemreX[7] among others, stand testament to quite some effort in the same application domain, While Saffron cannot be positioned relative to all of these here given the limited space, we now compare Saffron with the most related work: ArnetMiner[8] [6], a well-known state of the art "academic researcher social network search" tool.

ArnetMiner has an emphasis on classification and consists of two main parts. In the first part, probabilistic topic models such as Latent Dirichlet Allocation (LDA) [1] are extended and a unified topic model for papers, authors and conferences is proposed. It seems only the content of the papers is analyzed, not the calls for papers or the user interests from the homepages. They cluster all the words into 199 topics, which is a rather small number considering they analyze over a million papers (presumably from various fields). So, to make things clear, in ArnetMiner a topic means a group of words, not the name of a scientific area or technology as in Saffron. So then, whereas Saffron automatically extracts expertise topics, ArnetMiner does not. Instead, it classifies papers into 199 predefined research topics.

Furthermore, Saffron makes use of existing structured data on the Web, which is not addressed by ArnetMiner. The purpose of the other part of ArnetMiner is to find researcher profiles. Here profile means personal details extracted from homepages. They extend the FOAF ontology with other information (e.g education, research interest). They also deal with the name ambiguity problem. Saffron solves many of these problems through it's use and extension of the Semantic Web Dog Food corpus (structured information about researchers). This is enabled by:

1. The links from the SWDF corpus to more structured data about researchers.
2. The ease with which data crawled from these links can be merged with that already in SWDF.
3. The ability to consolidate the merged data due to the semantics bestowed on it by the ontologies used.

[6] `http://www.rkbexplorer.com/`
[7] `http://www.semrex.cn/`
[8] `http://arnetminer.org/`

So it is clear that the above **semantic information processing plays a central role in allowing Saffron to achieve things that alternative technologies cannot do as well, or at all.**

5 Web Interface

Saffron provides an attractive and functional Web interface for human users. The objective of the interface is to provide a system beyond a research prototype which delivers a rich user experience while exploring research collections. To this end, the explicit design process was put in place before any implementation work started. We worked over quick and schematic design sketches, user scenarios and brainstorm sessions towards more refined design wireframes and specifics of the interaction model.

The Saffron development process was iterative. Each iteration finished with a prototype of our most refined design concepts. The prototypes enabled us to use, test, and 'feel' the system. We involved a small number of users (researchers) to observe how the prototypes are used, what the confusing parts of the design are and to collect informal feedback.

Saffron is an exploration system where user goals are diversified (e.g. find publications, experts or expertise topics) and specified to a different degree (e.g. a user might have a specific expertise topic in mind or have only a rough idea about it). The number and variety of scenarios we chose to support created a challenge for designing a usable yet simple system.

One of the outputs of our design process and early testing are the design goals which underlined development of the most recent iteration of Saffron:

Simple and uniform interaction model. The only interaction technique we wanted to use is mouse point-and-click since it is the most ubiquitous one. We wanted all resources to be clickable and behave the same way once they are clicked. Furthermore, we wanted this behaviour to be uniform with what people are used to when they surf the Web and click on links. (To avoid confusion, we decided that each click leads to an entirely new customised page rather than modifying parts of the current page (e.g. clicking on a researcher might only change a pane with currently visible publications)).

Controlling visual complexity. We wanted to clearly distinguish between various resource types (i.e. people, expertise topics and publications) but at the same time show that all the resources can be interacted with the same way - i.e. through mouse clicks. Furthermore, we wanted to communicate visual structure of the UI with minimal use of graphical cues.

Fast responses. To encourage exploration we wanted the response times between user actions and UI updates to be short. To this end, Saffron uses six different indices and a caching system. Not all the data, however, is static and

can be indexed. Whenever the data has to be pulled in real-time and notable delay can occur (e.g. finding connections between researchers using SPARQL endpoints) a requirement was to display all available UI elements as soon as possible, indicate that some data is still loading, and extend UI once the loaded elements are available.

Support specific and non-specific user needs. Apart from supporting users with less specific goals (through exploration - i.e. moving from one resource to another) we also wanted to support users who know exactly what they are looking for. To this end, Saffron provides search functionality. Search expressions can contain arbitrary number of terms and phrases (e.g. "Semantic Web") combined with boolean operators (AND, OR, AND NOT), prepended (+plus and -minus, indicating that the entity should be either required or forbidden), or logical groups delimited by parentheses.

Personalisation. A requirement was to personalise user experience for a user who is logged in. Saffron can show connections, joint publications and mutual contacts between the user and expert being viewed, as seen in Figure 1. This is an important step beyond the traditional role of expert finding into that of expert contacting.

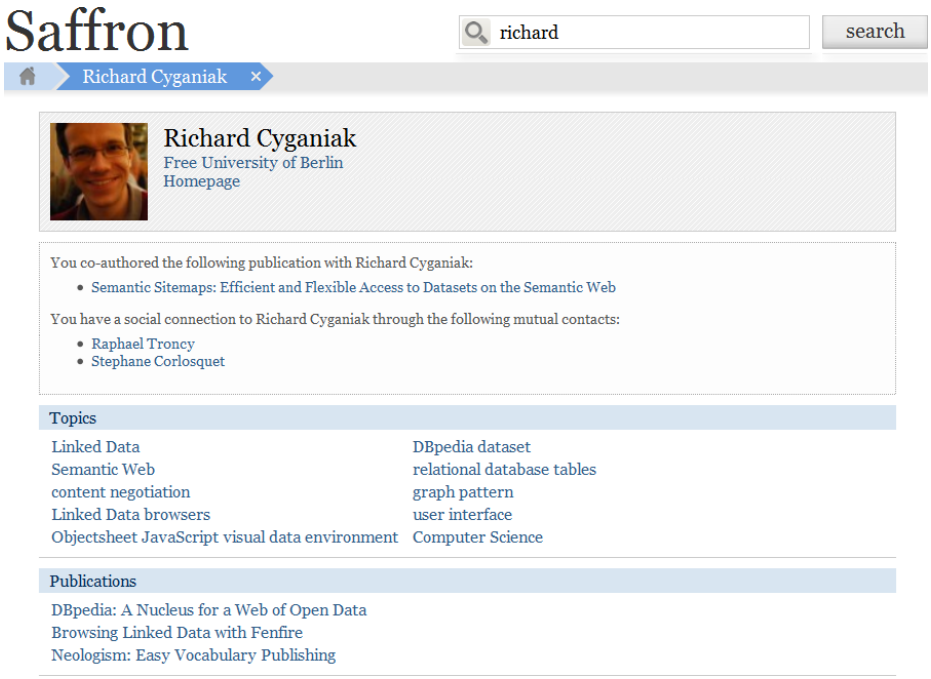

Fig. 1. Screenshot of user context used to personalise UI showing mutual contacts.

Do not rely on incomplete data. The LOD and SWDF contain valuable information but often it is incomplete or available only for a limited number of resources. Our requirement was to use the data to extend the UI views. And whenever data is not available, silently collapse blank spaces without breaking the visual structure of the information.

Most relevant information first. A requirement was to rank the visible pieces of information by showing the most important resources, in a given context, first.

Scalable in terms of the amount of data used. Note that our objective of delivering rich user experience did not involve UI design alone but rather had impact on all layers of the system. Therefore Saffron has been designed to be scalable in terms of the amount of data used. This included a requirement for indexing and caching components to enable quick system response.

Indices are built with KinoSearch implemented in C and wrapped in Perl. They scale to millions of entries providing stable and quick responses. On top of that we have a caching system which saves generated fragments of HTML. This infrastructure can scale to millions of documents, researchers and expertise topics without notable decrease in performance.

Acknowledgments. This work has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Líon-2).

References

1. D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 2003.
2. G. Bordea, P. Buitelaar, Deriunlp: A context based approach to automatic keyphrase extraction, in: SemEval 2010: Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation, 2010.
3. G. Bordea, P. Buitelaar, Expertise mining, in: AICS 2010: Proceedings of the 21st National Conference on Artificial Intelligence and Cognitive Science, 2010.
4. S. Kim, A. Medelyan, M. Kan, T. Baldwin, Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles, in: SemEval 2010: Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation, 2010.
5. F. Monaghan, Context-aware photograph annotation on the social Semantic Web, Ph.D. thesis, National University of Ireland, Galway (December 2008).
   URL `http://sw.deri.org/~ferg/publications/thesis.pdf`
6. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, Z. Su, Arnetminer: Extraction and mining of academic social networks.
7. G. Tummarello, R. Delbru, E. Oren, Sindice.com: weaving the open linked data, in: ISWC'07/ASWC'07: Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, Springer-Verlag, Berlin, Heidelberg, 2007.

Appendix: Summary of Address of Evaluation Criteria

The application has to be an end-user application, i.e. an application that provides a practical value to general Web users or, if this is not the case, at least to domain experts **(See Section 1, paragraph 1)**.

The information sources used **(See Section 3)**:

- should be under diverse ownership or control **(See Section 3, paragraph 1)**
- should be heterogeneous (syntactically, structurally, and semantically), and **(See Section 3, paragraph 4)**
- should contain substantial quantities of real world data (i.e. not toy examples). **(See Section 3, paragraph 5)**

The meaning of data has to play a central role **(See Section 4)**.

- Meaning must be represented using Semantic Web technologies. **(See Section 4, paragraph 1)**
- Data must be manipulated/processed in interesting ways to derive useful information and **(See Section 4, paragraph 4)**
- this semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all; **(See Section 4, paragraph 7)**

The application provides an attractive and functional Web interface (for human users) **(See Section 5, paragraph 1)**.

The application should be scalable (in terms of the amount of data used and in terms of distributed components working together). Ideally, the application should use all data that is currently published on the Semantic Web **(See Section 5, paragraph 12)**.

Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained **(See Section 2, paragraph 2)**.

Novelty, in applying semantic technology to a domain or task that have not been considered before **(See Section 4, paragraph 4)**

Functionality is different from or goes beyond pure information retrieval **(See Section 2, paragraph 1)**

The application has clear commercial potential and/or large existing user base

The application has clear commercial potential and has an existing user base in DERI, the largest Semantic Web research institute in the world. Ongoing and future work is also in the development of applications for larger organisations such as the National University of Ireland, Galway (NUIG) at large and in collaboration with industrial partners on the application of Saffron in corporate environments.

Contextual information is used for ratings or rankings **(See Section 2, paragraph 3 and Section 5, paragraph 9)**

The results should be as accurate as possible (e.g. use a ranking of results according to context) **(See Section 2, paragraph 3)**