# 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:** contextual information, Natural Language Processing, semantic information processing, Semantic Web Dog Food corpus, expertise topics, information retrieval, Semantic Web technologies, Linked Open Data[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 results 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] Saffron ate its own dog food by extracting these 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 of the many 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
- A researcher looking for publications on a specific topic

## 2 Expertise Topic Extraction

**Saffron's functionality is different from and goes beyond pure information retrieval.** A list of manually identified *skill types* (e.g. "algorithm") are used along with usage context to identify the *expertise topics* (e.g. "Machine Learning") that they introduce [3]. 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.

**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 assigned 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 judgement [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: this analysis measures the relation strength between a researcher and the expertise topics from its expertise profile by using the Sindice Semantic Web search engine[3] [7] (see Section 3). 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] `http://www.sindice.com/`

## 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[4] provides information on papers that were presented (including URL links to the source PDF files), 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 and the returned number of hits is considered as an additional 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 on the Linked Open Data (LOD) Web 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 (hereon crawl1 and crawl2 respectively), so that essentially two levels of depth from SWDF were crawled through. These include information from, amongst other sources, OntoWiki[5] as well as individual FOAF profiles and so 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 the one 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 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.

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

**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 on that subset of 747 papers that have such links to PDF files. Table 1(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. We analysed an average of 320 expertise topic candidates per document and an average of 142 expertise topic candidates per researcher. Furthermore, Table 1(b) gives numbers on the triple data crawled from the LOD Web, showing the total no. of triples, papers, people and social "knows" connections between people.

(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 |
| consolidated | 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, e.g. to build a holistic view of the social graph from incomplete data fragmented across sources that use different URIs for the same people. Additionally, SPARQL is used to query the data in an expressive way, e.g. to find indirect connections between researchers.

Furthermore, Saffron attempts to assign each extracted expertise topic to a concept URI from the LOD Web. In the current prototype we search for URIs from DBPedia. 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 and thereby the LOD Web.

**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. 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 analysed, and structured data, such as social connections, is not considered. 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 of structured LOD 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.
3. The ability to consolidate the merged data, e.g. repair broken social connections, due to the semantics bestowed on them by the ontologies used.
4. The ability to query the consolidated data expressively, e.g. to find indirect social/professional paths between researchers, due to the power of SPARQL.

The "open"-ness of LOD is crucial here to obtain a holistic view of the social graph: this would not be possible with alternative silos of such social information, such as Facebook or LinkedIn, since their APIs only provide access to one step into the social graph (the current user's contacts). When coupled with expertise topic extraction, this makes it clear that **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.** An objective of Saffron is to provide a system beyond a research prototype which delivers a rich user experience while exploring research collections.

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

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). Therefore, the number and variety of scenarios that must be supported creates a challenge for designing a usable yet simple system. To this end, an explicit design process was put in place before any implementation work started.

The Saffron development process was iterative. Quick and schematic design sketches, user scenarios and brainstorming sessions were used to move towards more refined design wireframes and specifics of the interaction model. Each iteration finished with a prototype of the most refined design concept. The prototypes enabled the system to be used, tested, and "felt". A small number of users were involved to observe how the prototypes were used, what the confusing parts of the design were and to collect informal feedback to improve the design of further iterations of Saffron.

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

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

**Controlling visual complexity.** While the various resource types should be graphically distinguished, it should be shown that they can be interacted with in the same way i.e. through mouse clicks. Furthermore, the visual structure of the UI should be communicated with minimal use of graphical cues.

**Fast responses.** To encourage exploration, response times between user actions and UI updates should be short (within milliseconds). To this end, Saffron uses six different indices and a caching system. However, not all the data is static and can be indexed. Whenever the data has to be pulled in real-time and a notable delay can occur (e.g. finding connections between researchers using SPARQL endpoints) a requirement is to display all available UI elements as soon as possible, indicate that some data is still loading, and extend the 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), support should also be given to users who know exactly what they are looking for. To this end, Saffron provides search functionality. Search expressions can contain arbitrary numbers of terms and phrases (e.g. "Semantic Web") combined with boolean operators (AND, OR and NOT). Terms and phrases can also be prepended with plus (+) or minus (-), indicating that the entity should be either required or forbidden respectively. Logical groups can be delimited by parentheses.

**Personalisation beyond pure information retrieval.** A requirement was to personalise the user experience. Saffron can show connections, joint publications and mutual contacts between the user and the 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.

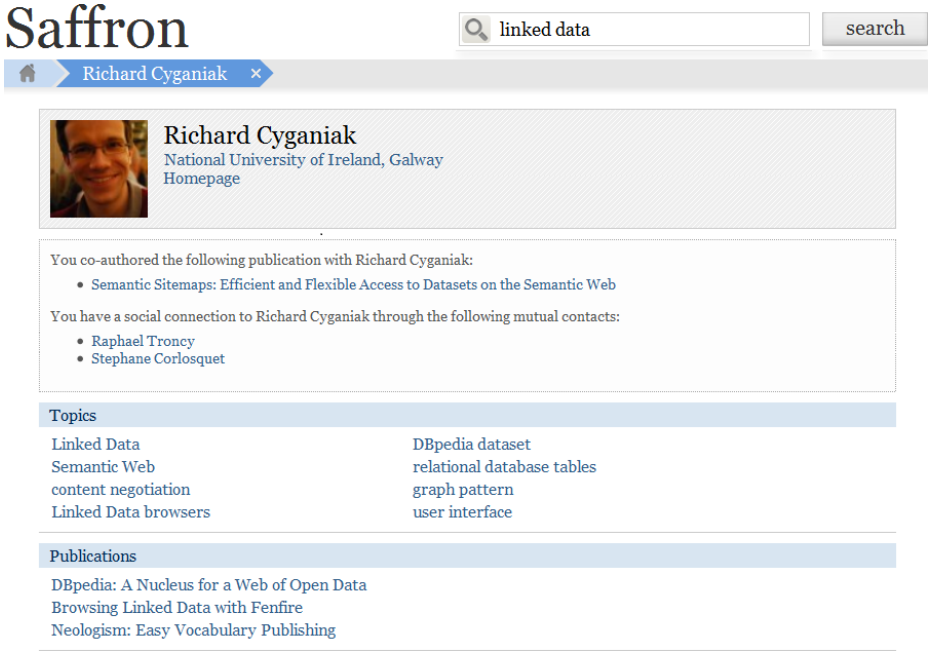


**Fig. 1.** Screenshot showing mutual contacts between login user and retrieved expert.

**Do not rely on incomplete data.** The LOD Web and SWDF contain valuable information but often it is incomplete or available only for a limited number of resources. Therefore, a requirement was to use the available data to extend the UI views. Whenever data is not available, blank spaces should be silently collapsed 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 the objective of delivering a 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 includes a requirement for indexing and caching components to enable quick system response.

Indices are built with KinoSearch[9], which is implemented in C and wrapped in Perl. It scales to millions of entries providing stable and quick responses. On top of that Saffron has a caching system which saves generated fragments of

[9] `http://www.rectangular.com/kinosearch/`

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: Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010), 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. N. Kim, A. Medelyan, M.-Y. Kan, T. Baldwin, SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles, in: Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010), 2010.
5. F. Monaghan, Context-aware photograph annotation on the social Semantic Web, Ph.D. thesis, National University of Ireland, Galway (December 2008).
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 Challenge Criteria

The application has to be an end-user application **(See Sec. 1, paragraph 1)**.
The information sources used **(See Section 3)**:
- should be under diverse ownership or control **(See Section 3, par. 1)**
- should be heterogeneous, and **(See Section 3, paragraph 4)**
- should contain substantial quantities of real world data. **(Sec. 3, par. 5)**

The meaning of data has to play a central role **(See Section 4)**.
- Meaning must be represented using Semantic Web technologies. **(S. 4 p. 1)**
- Data must be manipulated/processed in interesting ways to derive... **(S4p4)**
- this semantic information processing has to play a central role in... **(S4p7)**

The application provides an attractive and functional Web interface (for... **(S5p1)**.
The application should be scalable (in terms of the amount of data... **(S5p12)**.
Rigorous evaluations have taken place that demonstrate the benefits... **(S2p2)**.
Novelty, in applying semantic technology to a domain or task that have... **(S4p4)**
Functionality is different from or goes beyond pure IR **(S2, p1 and S5, p9)**
The application has clear commercial potential 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 in the application to larger organisations such as the National University of Ireland, Galway and, in collaboration with industrial partners, to corporate environments.**

Contextual information is used for ratings or rankings **(S. 2, p. 3 & S. 5, p. 9)**
The results should... (e.g. use a ranking of results according to context) **(S2p3)**