Exploiting Online Data in the Policy Making Process

Aron Larsson[1,2], Steve Taylor[3], Timo Wandhöfer[4], Vasilis Koulolias[1]

[1] eGovLab, Dept. of Computer and Systems Sciences, Stockholm University
[2] Dept. of Information and Communications Systems, Mid Sweden University
[3]IT Innovation, University of Southampton
[4]GESIS Leibniz Institute for the Social Sciences

aron@dsv.su.se, sjt@it-innovation.soton.ac.uk,
timo.wandhoefer@gesis.org, vasilis@egovlab.eu

Keywords: Policy making, policy analysis, social media, semantic web, open data, Sense4us, decision support

Abstract. This paper reports on the ambitions and methods behind the Sense4us project, aimed to provide ICT tools supporting policy making through systematic gathering of heterogeneous online data to increase problem understanding and the general public's opinions. The tools' goal is to enable stakeholders within the political sphere to identify online available data concerning their policies.

1 Introduction

Policy making is a complex activity since it involves striking a balance between legal requirements, intended outcomes and public response to the policy. Whilst incorporating popular input into the process is crucial to the legitimacy and acceptability of the outcome, it is also desirable to match citizen's expectations and demands to the policy. Questions of great concern for policy makers then become when policy makers are assured of that sufficient relevant issues and influences are taken into account, and to what extent the impact of a policy can be predicted before it is implemented. This is both in terms of the policy issue and targets themselves, i.e. will the effect of the policy reach what is desired, as well in terms of how the policy is accepted by citizens and stakeholders, which is important since the policy impact may be dependent on broad acceptance, avoiding non-desired outcomes and early needs for reformulation or abandoning the policy, cf., e.g., [8].

However, much of the literature on public policy analysis deals with (ex-post) evaluation, which tries to understand the causes and consequences of policies after they have been implemented [17]. Ex-ante evaluations are however equally important, carried out at the early stages of policy development and having a prescriptive bent involving impact assessment and ranking of policy options, see [18]. In this stage, citizens and policy makers alike wrestle with how to intelligently filter information according to relevance, relationship and provenance. Ex-ante evaluation encompasses forecasting

of consequences if policies were to be implemented and prescriptions about which policies should be implemented. One important aim of ex-ante decision support within the context of policy making therefore involve developing ways of facilitating for policy-makers to create policies that is consistent with their preferences while at the same time being accepted by other stakeholders, cf. [3]. The policy making challenges then also include sense making and trust building within the constraints of a participatory exercise – communicating the important issues and why there are strong beliefs in a certain policy while being aware of public opinion with respect to the issue. Decision makers are at the same time increasingly coming under pressure to be more inclusive and co-create policy with stakeholders, both from technologists as well as international and regional treaties such as the Aarhus Convention (1998). Recognising the importance of participatory practices in the network society implies looking not only at what happens in formal participatory practices, but also at what happens behind the scenes, in informal practices [4]. These informal practices are not necessarily organised in invited spaces, but are emerging spontaneously and are based on common concerns created by the particular situation at hand [6]. This relates particularly to the use of social media when framing policy decisions or anticipating their impact.

Previous attempts on providing ICT tools supporting this task has mainly focused on finding procedures for the incorporation of decision data obtained from decision makers and experts. Less work has been done on the means for providing information on both the public's views, values, and opinions without initiating directed polls, together with fast means for obtaining facts and knowledge about the policy issue at hand by searching for published datasets and reports. Traditional methods for gathering opinions are limited to polls, surveys and on-line portals, all of which are open to the biases which arise from the framing of questions and self-selection of respondents, also coming with the expensive need to design and adapt the means used for gathering the information. Additional efforts must also be put on the identification of relevant datasets, the finding of relevant reports, and understanding a complex network of stakeholders, all activities which could be effectively facilitated by novel methods for searching on-line data. In other words, on-line data can support basing a policy decision on both public opinion and "evidence" on that it will be effective, increasing the likelihood of broad public acceptance and that targets will be reached.

1.1 Online Data

Online data, i.e. data that can be accessed remotely and physically resides on a device connected to a wide area network, range from sensor data to text, from social media to expert repositories of knowledge. A vast ocean of heterogeneous information available online emerges, however sifting through this ocean and finding the data relevant for a policy issue at hand seems a task too difficult to overcome. A matter of great concern for contemporary technological development in the e-government domain is to what extent this difficulty can be remedied by systematic methods implemented in web tools, simple enough to be used for policy makers and analysts when they enter a new policy issue? As one contribution to this area, the Sense4us project is developing search tools to help find and present relevant sources of information and is building social media analytics tools to discover and track what people are talking about that is relevant to the topic of interest.

Of significance, the project is devoted to develop a software modelling tool that helps policy makers to assemble the information they have discovered and link it together. This enables the influences and impacts of policy to be investigated, and its likely outcomes identified. The ambition is therefore to provide aid to policy makers in their struggle to discover knowledge and opinions about the policy issue at hand, and this in turn helps them to capture perspectives that they would not normally be aware of or taken into account in the policy formulation stage of the policy making cycle. See [9] for a comprehensive treatment of this cycle.

2 The Toolkit

The toolkit revolves around online annotated data enabling for thematic searches using keywords and/or so-called hashtags. Two such data sources are in focus, namely social media data, currently focusing on Twitter feeds, and linked open data, i.e. data or datasets that are open and linked semantically, thereby the name of "semantic web" is often used when referring to linked open data.

2.1 Social media´

Recognising if a policy is well or badly received by the citizens, what elements of the policy are more controversial, and who are the citizens discussing about that policy are key factors to support policy makers in understanding, not only the citizen's opinions about a policy, but also up to which level the social media dialogs represent public opinion and should be used to inform the policy making process.

Following this purpose, the research and development of accurate sentiment analysis tools is at the core of the Sense4us project. We have investigated the use of contextual and conceptual semantics from Twitter posts for calculating sentiment [12] [13] [14]. This involved running a comparison of the two types of semantics with respect to their impact on sentiment analysis accuracy.

Results showed that using conceptual semantics (gleaned from term co-occurrence) improves sentiment accuracy over several baselines. Results also showed that adding conceptual semantics (entities extracted using AlchemyAPI) enhances this accuracy even further.

Accuracy is key in the context of Sense4us since the project aims to provide trustable information in which policy makers can support their decisions. Following this goal we also studied the role of stop words on sentiment analysis [11], showing that best results are achieved when using automatically generated dataset-specific set of stop words. Furthermore, we experimented with a new approach to automatically extend sentiment lexicons to render them more adaptable to domain change on social media, and generated and published a new gold-standard dataset for social media sentiment analysis [10].

2.2 Linked open data

The amount of semantic data published on the Web has increased considerably in the last years. One of the most remarkable efforts is the Linking Open Data community project1, which developed several tools and defined best practises for various steps of the semantic data lifecycle. More specifically, the project focused on creating, integrating, publishing, documenting, and validating so-called Linked Data (i.e. data that follows the Linked Data principles). As a result, the "Linked Open Data cloud" was created. The LOD cloud is a set of RDF2 datasets interlinked with each other, containing as of August 2014 datasets about several topical domains such as media, life sciences, government, publications, linguistic resources and social networking.

A central task of the project in the future is the improved accessibility of the Linked Open Data cloud for policy makers as well as project partners. The two major elements here are the ranking of available data sets within the Linked Open Data cloud, in regard to given queries. Both elements support policy makers to gain deeper insight in topics as well as providing them with additional information about related fields and possible effects of a policy.

The following **Error! Reference source not found.** displays the conceptual view how the "Policy Maker" benefits from published linked open data. The policy maker (see left) interacts with the Sense4us user interface (see green box in the background called "User Interface"). He has two options for retrieving the Linked Open Data Cloud. The outcome of both retrieval strategies is a ranked list of data sets in terms of relevance for the policy theme.

---

1 http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData,

2 Resource Description Framework, a web standard for data interchange.

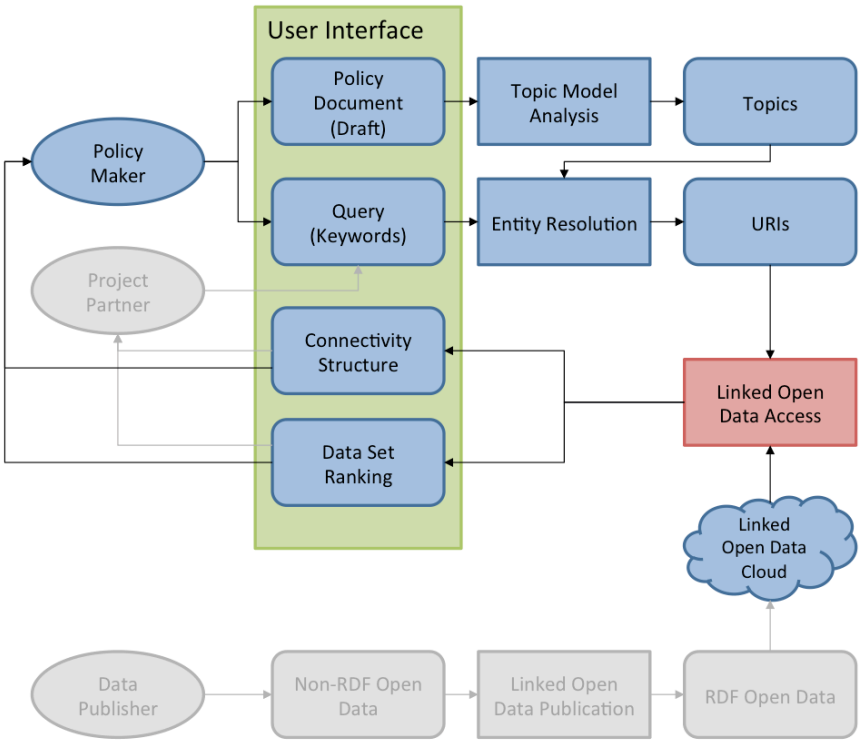

Fig. 1. Exploiting linked open data from a policy maker's perspective

2.3 Problem structuring

Problem structuring are concerned with facilitating policy evaluation, i) to measure the effects of a policy, or impact assessment, ii) to understand why the effects are to be, and iii) to facilitate learning about the policy issue at hand, cf, e.g., [16]. The prescriptive impact assessment is a challenge, where the effects of a policy are often delayed in time as well as characterized by multiple perspectives, conflicting interests, or uncertainties. To answer these challenges, problem structuring methods have emerged, aimed at facilitating to obtain a better understanding of unstructured problems. The methods rely heavily on engaging with policy makers, adopting a facilitative mode of engagement, and simple, often qualitative models [7].

Providing an ICT tool for problem structuring tailored for modelling of public policy problems involving entities such as policy instruments, goals and targets, and actors, where there is an underlying causal map representation of how changes in instruments lead to change in goal variables. See [1] for a detailed presentation of causal maps. It is possible to simulate policy consequences and possible future scenarios on the causal map by quantifying elements of the map (variables and change transfer coefficients). Further, scenarios, or alternative policy options based on a forward-looking impact as-

sessment in terms of economic, social, environmental and other impacts can be generated and decision evaluation of the generated options can be done with decision analysis methods.

Scenario generation helps policy-makers in identifying feasible options from a possibly vast space of possible ones reaching stipulated targets, while the decision evaluation can supports an in-depth performance evaluation of policy proposals taking the preferences of actors into account. The aim is to provide a policy-oriented software solution that implements a systems approach to structure a public policy problem situation and simulate the system behaviour and responses to interventions over time using a dynamic simulation model, in order to design policy options and assess the consequences given a number of alternative possible futures. Finally, model building also requires access to large amounts of information and means for identifying the elements of the problem model, which is often a constraint for modelling activities. In this respect, it is of high concern to investigate the interface between fast web based means for gathering and filtering policy relevant information, such as linked open data searches and sentiment analysis, in order to facilitate efficient use of a problem structuring tool.

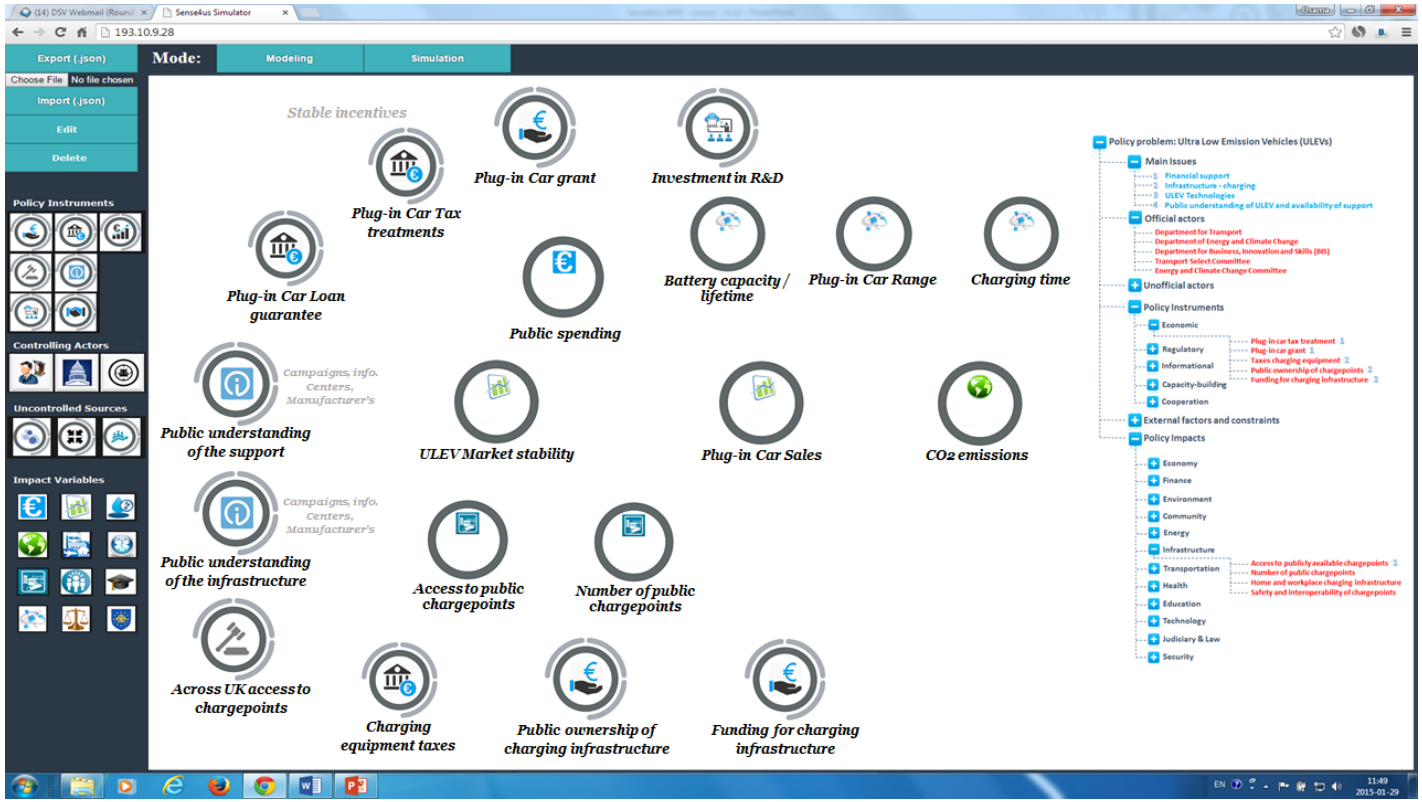

Fig. 2. Problem structuring with causal maps.

3 The Sense4us Platform as an Integrated Toolkit

The toolkit is an integrated framework that enables the user to use information gathering and analysis tools to address the informational challenges described above. The tools are in the following areas.

- A text summarisation tool enables the user to find major key words or phrases in documents or other bodies of texts (such as a document or collection of social media postings). The tool uses the well-established Latent Dirichlet Allocation (LDA) technique described by Blei et al. [2] and applied to social media in [15]. The benefit to the user is that they can determine the key themes of the texts without reading them, enabling them to prioritize which texts should be read first.
- Finding related information – given a key word or phrase, this tool enables the user to find information around the key word's theme in open data or social media, and how the information is related. We are currently investigating automated searches of DBpedia[3] (which is a semantically annotated version of Wikipedia). The user specifies a Wikipedia page, and the result of the search is a map of links of different types in and out the page. The benefit to the user is that they are able to find previously unknown information around their policy area of interest, thus increasing their body of knowledge about the policy area.
- Opinion analysis [5] [10] enables the user to discover what peoples' opinions towards the topic of interest on social media. Given a body of social media postings, the tool can indicate the overall sentiment towards key words of interest to the user.
- Policy modelling and simulation [1] enables the user to create a model of their policy (using the other tools to provide information for the model). This model takes the form of a desired outcome and some policy options that may achieve that outcome. The options may be evaluated via simulation using group decision and negotiation analysis techniques to examine the impact of the policy option on the desired outcome and different classes of citizen (in effect who wins and who loses).

Each tool can be used as and when the user needs it, and some tools' output may also be used as the input for another tool, enabling the user to gain deeper insight by additional analysis on data. The overall system architecture is shown in **Error! Reference source not found.**.

[3] http://wiki.dbpedia.org/

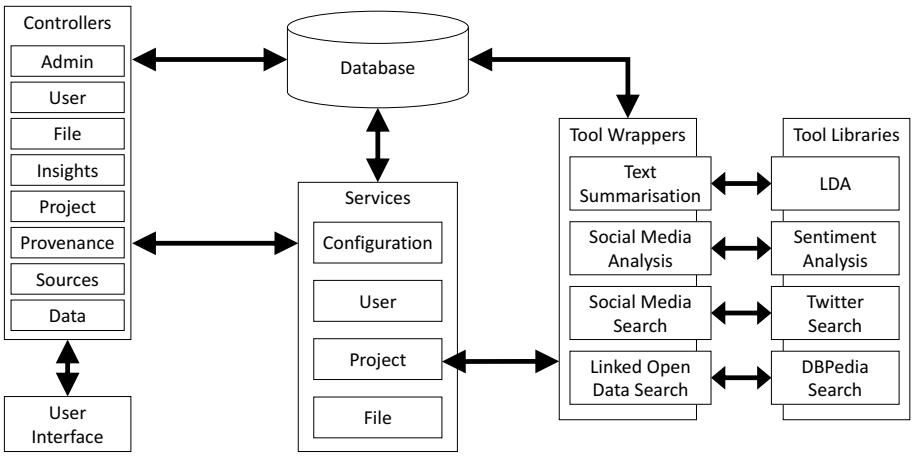

Fig. 3. System Architecture

The system uses the Spring Framework4, an application framework for hosting Java applications. This utilises a "model-view-controller" approach, which separates the data model from the management logic and the presentation layer. This was chosen to allow a significant amount of flexibility in how the toolkit is controlled by the user and how the results are presented. The presentation layer is named "controllers", and these are responsible for displaying results and acquiring control input from users. Management logic is represented by "services", which contain control logic and processing beyond the capability of controllers.

All data input into the system or output from tools is stored in a database, together with provenance information, which includes a description of the data source, the time and date the data was processed (collected or transformed by a tool), and which user owns the data. The database is MongoDB5, a so-called "NoSQL" database, chosen for its flexibility in storing semi-structured data such as JSON, which is the output of many of the data sources and tools in the toolkit.

In order to enable tools to be added as necessary, each tool is interfaced to the rest of the system by wrappers, and examples are shown in **Error! Reference source not found.**. The wrappers perform functions such as fetching input data from the database, preparing it for the tool (e.g. formatting as necessary), running the tool and storing the output into the database with associated provenance information. A new tool can be added to the toolkit by creating a new wrapper for it. In practice, this often means making a copy of the closest existing wrapper, and modifying the copy as necessary to create a new wrapper.

[4] https://spring.io/
[55] https://www.mongodb.org/

4 Summary

In this paper we presented information and communication technologies that are part of the research project Sense4us. Concerning the research regarding the semantics from Twitter posts we have investigated the use of contextual and conceptual semantics for calculating sentiment. Results showed that using conceptual semantics (e.g. gleaned from term co-occurrence or entities extraction using AlchemyAPI) the sentiment accuracy could be increased over several baselines. Regarding the conceptual semantics we looked at stop words where the best results are achieved when using automatically generated dataset-specific set of stop words. Furthermore, we experimented with a new approach to automatically extend sentiment lexicons to render them more adaptable to domain change on social media, and generated and published a new gold-standard dataset for social media sentiment analysis.

With respect to linked open data, since open data is provided in a variety of different portals, interfaces and formats on the web, we must advise data publishers of particular data sets how they can transform and publish their originally non-RDF open data in RDF format. When facing vocabulary design of data that is to be transformed, best practices of the Linked Data community should be followed like the reuse of existing vocabularies as much as possible. Additionally, an extensive data publication in RDF allows for detecting more suitable entities for the interlinking process.

The proposed policy modelling and simulation approach allows simplifying and summarising the decision maker's knowledge, notions, and causal beliefs, as well as information gathered from different sources about a social, socioeconomic or sociotechnical system.

References

1. Acar, W.; Druckenmiller, D., 2006: Endowing cognitive mapping with computational properties for strategic analysis, Futures 38:993-1009.
2. Blei, D. M.; Andrew Y. Ng, Michael I. J., 2003: Latent dirichlet allocation, Journal of Machine Learning Research 3: 993-1022.
3. Bryson, J. M. 2007. What to do when stakeholders matter, Public Management Review 6(1):21-53.
4. Cornwall, A., 2002: Making spaces, changing places: Situating participation in development. Institution of Development Studies, IDS Working Paper 170.
5. Fernandez, M.; Wandöfer, T.; Allen, B.; Elisabeth Cano, A.; Alani, H., 2014: Using Social Media To Inform Policy Making: To whom are we listening?. In Proceedings of the European Conference on Social Media (ECSM). UK.
6. Fung, A., 2006: Varieties of Participation in Complex Governance, Public Administration Review, vol. 66, 2006, pp. 66-75.
7. Franco, L. A.; Montibeller, G. (2010). Facilitated modelling in operational research. European Journal of Operational Research 205: 489-500.
8. Haggett, C, 2011: Understanding public responses to offshore wind power, Energy Policy 39(2): 503-510.
9. Lindblom, C., 1968: The Policy-making Process, Prentice-Hall, Englewood Cliffs NJ.
10. Saif, H.; Fernandez, M.; He, Y.; Alani, H., 2013: Evaluation datasets for twitter sentiment analysis a survey and a new dataset, the sts-gold. In Proceedings, 1st Workshop

on Emotion and Sentiment in Social and Expressive Media (ESSEM) in conjunction with AI*IA Conference, Turin, Italy, 2013.
11. Saif, H.; Fernandez, M.; He, Y.; Alani, H., 2014a: On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In Proc. 9th Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2014.
12. Saif, H.; Fernandez, M.; He, Y.; Alani, H., 2014b: SentiCircles for Contextual and Conceptual Semantic Sentiment analysis of Twitter. Extended Semantic Web Conference (ESWC), Crete, 2014.
13. Saif, H.; Fernandez, M.; He, Y.; Alani, H., 2014c: Adapting Sentiment Lexicons using Contextual Semantics for Twitter Sentiment Analysis. In Proceeding of the first semantic sentiment analysis workshop: conjunction with the eleventh Extended Semantic Web conference (ESWC). Crete, Greece.
14. Saif, H.; Fernandez, M.; He, Y.; Alani, H., 2014d: Semantic Patterns for Sentiment Analysis of Twitter, The 13th International Semantic Web Conference (ISWC), Riva del Garda - Trentino Italy.
15. Sizov, S. 2010; Geofolk: Latent spatial semantics in web 2.0 social media, Proceedings of the third ACM international conference on Web search and data mining. ACM.
16. Solliec-Berriet, M; Labarthe, P; Laurent, C, 2014: Goals of evaluation and types of evidence. Evaluation, 20(2), 195-213.
17. Tsoukiàs, A., Montibeller, G., Lucertini, G., and Belton V. 2013. Policy analytics: an agenda for research and practice, EURO Journal of Decision Processes 1:115–134.
18. Turnpenny, J.; Radaelli, C. M.; Jordan, A.; Jacob, K. 2009. The policy and politics of policy appraisal: emerging trends and new directions. Journal of European Public Policy 16(4):640-653.