Data Science for Public Policy

Of the people, for the people, by the people 2.0?

Sreenivas R. Sukumar and Mallikarjun Shankar
Computational Science and Engineering Division
Oak Ridge National Laboratory
1 Bethel Valley Road, Oak Ridge, TN, 37831, USA
Email: {sukumarsr@ornl.gov; shankarm@ornl.gov}

ABSTRACT

In this paper, we explore the role of data science in the public policy lifecycle. We posit policy documents (bills, acts, regulations and directives) as forms of social objects and present a methodology to understand interactions between prior context in professional and personal social networks to a given public policy document release. We employ natural language processing tools along with recent advances in semantic reasoning to formulate document-level proximity metrics which we use to predict the relevance (and impact) of the policy artifacts. These metrics serve as a measure of "excitation" between people and the public policy initiatives.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.3 [**Information Search and Retrieval**]: Information filtering, Relevance feedback

General Terms

Algorithms, Design, Human Factors

Keywords

public policy, social computing, semantic reasoning, relevance and impact estimation, data science, social good

1. INTRODUCTION

This paper is motivated by the potential of applying social computing to the public policy lifecycle in the context of Abraham Lincoln's vision in his Gettysburg address [1]. We illustrate the Lincoln vision of a government that is "of the people, for the people and by the people" in Figure 1. In this generally accepted model of public policy and governance, people contribute votes, money, and ideas through their chosen representatives. The government that is formed from such representation initiates, formulates, and prioritizes policy ideas to serve the people they represent. This typically happens by enacting policies that serve as law, regulations, and guidelines aimed at enhancing professional, personal, and social interactions in society.

Today, social computing (via different social media such as YouTube, Facebook, Twitter, Blogs and interactive web publishing) is revolutionizing the public policy lifecycle. Policy makers and politicians in the United States have embraced the digital media culture for advertising their leadership and communicating their governance manifestoes [2]. These advertisements succeed as social objects to influence people and determine who represents them in the government. In addition, governments are adopting social media to communicate directly with the people. The recent White House release [3] on transparent, accountable and collaborative governance publishes policy documents as readily accessible objects through which the government hopes to rapidly disseminate policy ideas, solicit access to expertise outside of the government, and identify opportunities for co-operation in policy making.

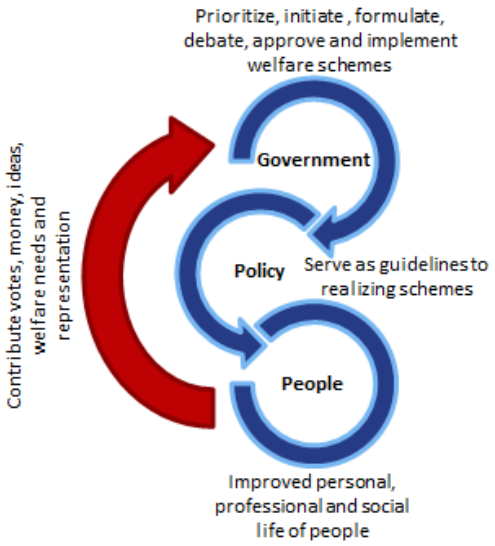

Figure 1. An illustration of the public policy lifecycle that tries to capture the Lincoln vision of a government that is of the people, for the people and by the people.

Recent (circa, spring and summer 2011) events in Tunisia, Egypt, Libya, and the U.K., have shown that social networks such as Facebook and Twitter dramatically enabled the citizenry to express their views. Social networks acted as tools to propagate ideas and mobilize movements leading in many cases to dramatic changes to the country's governance (as in Egypt), and in some cases undesired consequences (as in the UK). Clearly this is a shift from communicating through traditional media (newspaper, television, radio, etc.) to a kind of personalized expressive democracy – a mode of expression that governments should understand, listen to, and act upon to be effective and efficient. So, how can governments listen to people through social networks? How can social computing assist both the government and the people to improve the process of public policy making? Can we accept social networks as a quick census mechanism? Can we apply algorithmic thinking to conduct social science experiments as computer simulations and thereby predict

relevance and impact of a public policy idea? How can we use social computing to seek reliable feedback from the society?

As an attempt to find answers to the aforementioned questions, we have begun developing computational models and tools with specific application to public policy artifacts. We discuss a new tool here that is designed to: (i) evaluate policy consistency over time using semantic reasoning concepts, (ii) summarize and cluster expertise (skillsets) in the society, and (iii) estimate/visualize the interaction and societal responses to a policy publication. There is tremendous interest in evaluating social networks for a variety of purposes and we call attention to these efforts in section 2. In Section 3 we discuss our approach and describe the model we have developed that treats policy documents as social objects and evaluates how the policy objects excite representations (and opinions) of the individuals in society. We conclude with a summary in Section 4.

2. RELATED WORK

The idea of introducing Web 2.0 technologies into governance has garnered support from the software industry and policy makers. Software companies such as Microsoft [4] and SocialText [5] market Government 2.0 as an attempt to introduce enterprise social computing platforms into government organizations and operations. The underlying motive is that social networks can be repurposed to sharing and creating ideas across the public policy process.

Our work is oriented more towards the design of intelligent tools that automatically sense, understand, interpret and extract feedback from interactions within the objects present in such social platforms. In this sense it is closer to work in Tweetfeel [6] where the site is able to query twitter feeds for a keyword and after complex sentient analysis on tweet interactions can summarize the polarization of the people towards the keyword. For example, a search on President Obama in [6] reveals his popularity and acceptance amongst people that is at times surprisingly close to the estimate from meticulous surveys conducted by the news channels. In a similar fashion IBOPE/Zogby [7] is able to analyze interactive media such as blog posts and functions as a polling engine for public policy related ideas. Both [6] and [7] are proof of how social objects can be used as a census/survey mechanism. In their current form, though, these sites do not accommodate the wide availability of information related to an individual's expertise, credibility and strength of relationship with respect to a social network.

Our approach tries to also use such information to interface public policy documents to an individual based on interests and relevance. Several web-based software services are already able to extract interest profiles of people. MIT Persona [8] summarizes the web-presence of an individual along dimensions such as art, education, sports, etc. LinkedIn [9] publishes personal and career-related information that subscribers volunteer for public view. Microsoft Academic Search [10] is another social discovery tool that maps and links authors and collaborations based on authorship in journals and conferences. We consider these as objects that signify a profile (or profile element) that is "excited" by a policy artifact. We marshal such web-based services to sampling social object networks and extract the interest profiles for our work described in the following Section.

3. OUR APPROACH

There are two main components to determine how a policy artifact interacts with people and their social network as illustrated in Figure 2. The first component is a model that can extract the individual's interest profile embedded in their social media presence – profiles, posts, links, etc. The second component is a

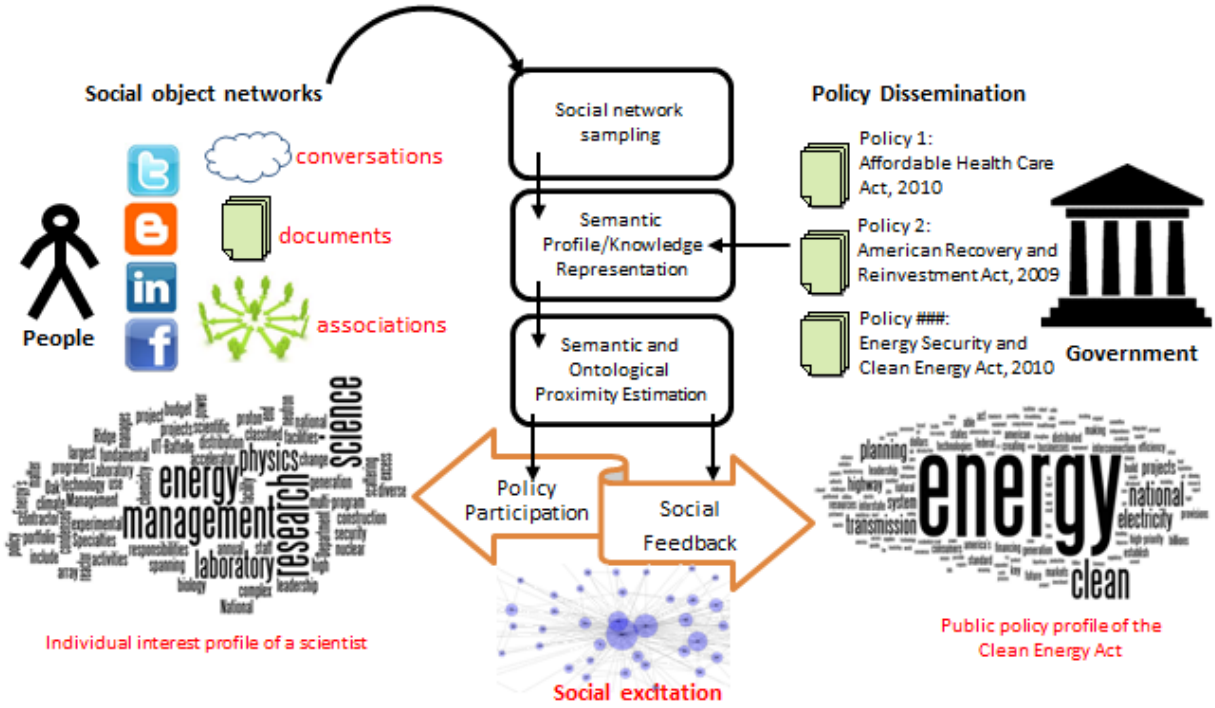

Figure 2. Data-flow in our proposed semantic model that links social objects and public policy.

systematic way to characterize how a document excites the models for the individuals and the social network components. We accomplish this by first extracting a socio-semantic representation using the profiles and prior interactions of individuals. Next, through a semantic mapping process, we estimate the relevance of the public policy social object (a government released document) to the individual (and his or her network). The semantic mapping process uses semantic-proximity metrics that quantify how much a policy-related social object is expected to excite the whole or a part of a person's representation.

We develop the socio-semantic representation by incorporating profiles from publicly shared social object publications in LinkedIn, personal websites, etc. The interest profile is an ordered list of keywords that emerge as salient from their interactions – this creates the basis for adding semantic value to the interest profile. We have visualized an example as a tag cloud [11] in Figure 2. The size of the word in the tag cloud encodes the saliency of the word – large font size encodes higher saliency. We use the term frequency- inverse document frequency (TF-IDF) formulation [12] from the natural language processing literature to extract saliency from the text content available in social networks. All the tweets, blog posts, LinkedIn and Facebook profiles serve as documents associated with an individual's profile to compute 'term' frequency. The collection of such documents over a social network constitutes our corpus while computing the TF-IDF values. We use the TF-IDF value as weights for the words that describe the individual's expertise and interests. The collection of lists of salient words and weights over a social network is the simplest form of the knowledge representation of the individuals and their networks. Similarly, and somewhat symmetrically, we extract a profile from the policy document.

We show the data flow between social object networks, the policy objects and our computational engine in Figure 2. The example chosen for the individual is a scientist at the Oak Ridge National Laboratory and the corresponding policy document chosen was the Energy Security and Clean Energy Act of 2009. In the following paragraphs, we detail how to compute the semantic proximity between the two profiles and discuss experimental results that show our ability to use these social measures as an indicator of social excitation.

3.1 Extracting interest and policy profiles

The social network sampling and the keyword-based profile extraction process result in two weight-associated vectors of words as shown in Figure 3.

Interest profile from a social object: *s*

w_1	w_2	w_3			$w_{Nperson}$
Energy	Research	Science	….	…..	Technology
1	2	3			N_{person}

Policy profile: *p*

q_1	q_2	q_3			$q_{Npolicy}$
Energy	Clean	Planning	….	…..	National
1	2	3			N_{policy}

Figure 3. Interest profile extracted from the individuals interactions in social platforms. An example of a profile (*s*) extracted from interactions on LinkedIn by one of the authors working at the Oak Ridge National Lab and the policy profile (*p*) for the Energy Security and Clean Energy Act 2009.

The number of salient words used to represent the profiles of individuals (N_{person}) and the policy document (N_{policy}) are currently user-specified parameters. The values for these two parameters are chosen as a tradeoff between speed and the desire for better results. The weights w_i and q_j are the TF-IDF values.

3.2 Computing semantic proximity between social interest and policy profiles

We begin with the semantic-proximity measure. The domain agnostic semantic-proximity measure uses a special organization of words in *WordNet*® [13, 14][1]. The complex *WordNet* is often best visualized as a dictionary of words that is organized as a tree based on the conceptual-semantic headers. Within this tree, we search for all possible paths between a word in the individual's interest profile s_i and another word in the policy profile p_j. We count the number of hops $h(s_i, p_j)$ that one has to make to link the two words within the *WordNet* tree structure. We illustrate this idea in Figure 4. Some pairs of words can result in multiple paths of tree-traversal. In such cases, we take the average number of hops over all possible paths. We repeat this for every word pair that includes one word from the interest profile and another from the policy profile. We repeat this process for every individual in the sampled social network.

For the k^{th} individual in the network, the semantic proximity measure (SPM) can be defined as Equation 1.

$$SPM(k) = \sum_{i=1}^{N_{personal}} \sum_{j=1}^{N_{policy}} \left[w_i * q_j * h(S_i, P_j) \right] \qquad (1)$$

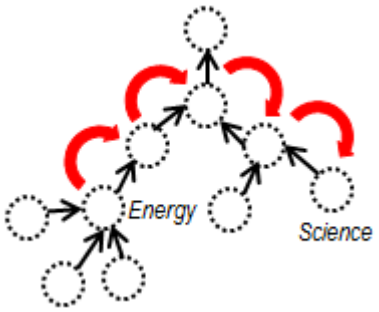

Figure 4. Semantic proximity in the *WordNet* Tree.

The *SPM* is a domain agnostic measure and does not consider context in which words come together. The ontological proximity measure considers such aspects but requires a multi-dimensional tree search. We illustrate the idea in Figure 5. The ontological layers/dimensions in the figure are 'Deploys', 'Located at', 'Funds', 'owns' etc. The inclusion of layered ontologies introduces context into the relevance measures. We attribute weight vectors v_l for the l layers of attributes within the ontology. Verisimilar to how we computed the *SPM* for each word pair (s_i, p_j), we march through the multi-dimensional tree counting the number of ontological hops $g(s_i, p_j)$ required to traverse from one word in the individuals interest profile *s* and another in the policy profile *p*. We compute the ontological proximity metric using Equation 5 for the k^{th} individual in the social network.

[1] *WordNet*® is a large lexical database of English in which nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms each expressing a distinct concept. The cognitive synonyms are interlinked by means of conceptual-semantic and lexical relations.

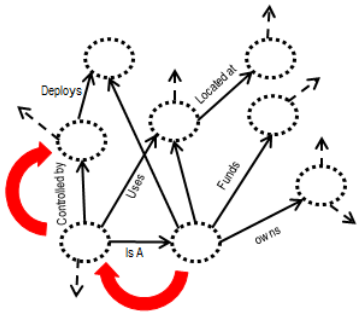

Figure 5. Computing the ontological proximity measure by hopping through multiple layers of domain specific definitions.

$$OPM(k) = \sum_{i=1}^{N_{personal}} \sum_{j=1}^{N_{policy}} \left[w_i * q_j * g(s_i, p_j) * \sum_{m=1}^{g(i,j)} v_{i,j,m} \right] \quad (2)$$

The relevance metric is then the sum of SPM and OPM.

$$R(k) = SPM(k) + OPM(k) \quad (3)$$

Both SPM and OPM measures are normalized for a network and a corpus of interest profiles before computing the relevance metric *R*. Our posit is that relevance measure *R* ranks the individuals whose presence was sampled from the social networks in order of relevance to the policy profile. Although, in future we would like to incorporate public policy ontologies similar to the ones described in [15], the results we present in the following paragraphs leverages the ontologies in Divisi [16] and ConceptNet [17,18].

3.3 Experiments and Results

In this section, we present results from preliminary testing of the semantic proximity metrics. We sampled an expertise network of people from an academic engineering community of 150 authors. We created their interest profiles from their articles in journals and conference proceedings and estimated the excitation of such a network with respect to excitations in the form of documents both within the expertise domain of the network and outside. Our excitation documents in this study were: (i) a research article from the same domain as the author's expertise, (ii) a research article from an auxiliary scientific community related to the domain, (iii) a sports news article and (iv) a movie recap from a blog post. The results are presented below in Table 1. We computed the proximity metrics presented earlier in this section and normalized the scores over the entire network between 0 and 1 such that higher values signify greater excitation. We present the mean of the normalized relevance measure in Table 1 for each of these excitations. We also tabulate the fraction of the sampled population that was excited by each of the target documents in Table 1.

Table 1. Excitation of the social network to policy documents

	Theme				
	Related to the domain expertise		**Outside the domain expertise of the network**		
	Domain	*Auxiliary*	*Sports*	*Movies*	*Travel*
% excited	97	76	6	6	1
Average R	0.71	0.62	0.18	0.11	0.04

We defined excitation as the deviation beyond one standard deviation of the average relevance measure from a profile to a corpus consisting of several thousand documents. The results show that our proximity metric is indeed topic-sensitive. It shows the excitation amongst people in a social network when the topics are of relevance while intelligently ignoring irrelevant content. The novelty in our approach is the ability to estimate relevance with respect to the social network of profiles in addition to the personalized interests of the individual.

4. SUMMARY AND FUTURE WORK

We have presented preliminary results from a model associating an individual's interest profile embedded in their social artifacts and objects – profiles, posts, links, etc., with policy social objects such as documents and web-based discussions. We have formulated a systematic way to characterize and compute how a document excites the models for the individuals based on their interests and the influence of the social network on their interests. We accomplished this by first extracting a socio-semantic representation using the profiles and prior interactions of individuals and through a semantic mapping process. We presented semantic-proximity metrics that are able to quantify how much a policy-related social object is expected to excite the whole or a part of a person's representation. We believe such a model is a fundamental first step in being able to link social computing and public policy.

The advantage of using the proposed model is that it can cater to the personalized delivery and interaction with the people and stimulate participation in the public policy lifecycle. The data and the models also can serve as a feedback mechanism for the government agencies. The semantic distribution of interests and expertise at a social, professional and personal level would allow the government to understand the needs of the society better. For example, understanding the skillset of a geographic region can be a critical to making investments that can create jobs in that area. Our vision is to emulate the success of the commercial interests in exploiting social networks to learn about people's interests for targeted advertisements and product surveys to the domain of public policy.

5. ACKNOWLEDGMENTS

This manuscript is authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

6. REFERENCES

[1] G. Wills, Lincoln at Gettysburg: The Words That Remade America. Touchstone Books, 1993.

[2] D. Talbot, "How Obama Really Did It: The social-networking strategy that took an obscure senator to the doors of the White House". MIT Technology Review, 111(5), 78-83, 2008.

[3] http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment/ ; Accessed 17 August 2011.

[4] http://www.microsoft.com/industry/government/guides/OpenGovernment/default.aspx; Accessed 17 August 2011.

[5] http://www.socialtext.com/; Accessed 17 August 2011.

[6] http://www.tweetfeel.com/; Aecssed 17 August 2011.

[7] https://zogby.com ; Accessed 17 August 2011.

[8] http://personas.media.mit.edu/; Accessed 17 August 2011.

[9] http://linkedin.com ; Accessed 17 August 2011.

[10] http://academic.research.microsoft.com/; Accessed 17 August 2011.

[11] F. B. Viegas, M. Wattenberg, J. Feinberg, "Participatory Visualization with Wordle," IEEE Transactions on Visualization and Computer Graphics, pp. 1137-1144, November/December, 2009.

[12] G.Salton, and C. Buckley, "Term-weighting approaches in automatic text retrieval", Journal of Information Processing and management, 24(5): 513-523, 1988.

[13] G. Miller, "WordNet:A Lexical Database for English," Comm. ACM, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[14] http://wordnet.princeton.edu/perl/webwn?s=word-you-want; Accessed 17 August 2011.

[15] E. N. Loukis, "An ontology for G2G collaboration in public policy making, implementation and evaluation", *Journal of Artificial Intelligence and Law*, 15, 1 (March 2007), 19-48, 2007.

[16] E. Cambria, A. Hussain, C. Havasi, C. Eckl, Common Sense Computing: From the Society of Mind to Digital Intuition and Beyond. In: Fierrez, J. et al. (eds.) BioID MultiComm 2009. LNCS, vol. 5707, pp. 260–267. Springer–Verlag, Berlin Heidelberg, 2009.

[17] H. Liu and P. Singh, "ConceptNet: A Practical CommonSense Reasoning Toolkit," BT Technology J., vol. 22, no.4, 2004, pp.211–226.

[18] C. Havasi, R. Speer, and J. Alonso,"ConceptNet 3: A Flexible, Multilingual Semantic Network for Common Sense Knowledge," Recent Advances in Natural Language Processing (RANLP 07), John Benjamins, 2007.