Bringing Citizens' Opinions to Members of Parliament

The Newspaper Story

Ruxandra Geana*, Steve Taylor**, Timo Wandhoefer***

** University of Southampton, UK, geana.ruxandra@gmail.com*

*** IT Innovation Centre, University of Southampton, UK, sjt@it-innovation.soton.ac.uk (contact author)*

**** GESIS – Leibniz Institute for the Social Sciences, Germany, timo.wandhoefer@gesis.org*

Abstract: We describe a method whereby a governmental policy maker can discover where their policy statements are discussed, and we show some example results for a case study validating our approach. Our strategy is to find news articles pertaining to the policy statements, then to perform internet searches for references to the news articles' headlines and URLs. We have created a software tool that schedules repeating Google searches for the news articles and collects the results in a database, enabling the user to aggregate and analyse them to produce ranked tables of sites that reference the news articles. Using data mining techniques we can analyse data so that resultant ranking reflects an overall aggregate score, taking into account multiple datasets. We can also examine differences between datasets, for example how the sites where the article is discussed change over time.

Keywords: e-participation, e-government, social networking sites, news articles, news stories, political engagement, citizens' opinions, dialogue

Acknowledgement: The WeGov project (no. 248512) is funded with support from the European Commission under the SEVENTH FRAMEWORK PROGRAMME THEME ICT 2009.7.3 ICT for Governance and Policy Modelling.

The work reported here has been done within the context of the WeGov IST FP7 project. The project's primary remit is to enable effective dialogue and engagement between e-governments and citizens, and a key feature of the project is that it uses social networking sites (SNS) as the primary communication channel. Before the project, there were a number of efforts to engage citizens with governmental policy, mainly using bespoke websites whose main drawback was that they were rarely used (see Hansard Society, 2009 for an example). WeGov is aiming to address this drawback by using tools the citizens already use: social networking sites, blogs, forums, etc.

The project supports its target audience of governmental policy makers with tools to enrich the two-way dialogue with citizens on SNS. The project's philosophy has been to develop a set of tools enabling the user to find and analyse SNS postings, and to make responses into SNS; along with a dashboard-based environment where the tools can be used individually or together.

We have adopted a methodical approach for the development process of the software with frequent and iterative end user engagement so as to get requirements and feedback on development progress. As part of user engagement, a number of use cases were designed showing how the WeGov analysis tools could provide a two-way dialogue with citizens (see Addis et al, 2010), and the work reported here develops one of these use cases.

An important aspect of the work in WeGov is to protect the rights and privacy of citizens and policy makers. To address this, a legal and ethical analysis was conducted to provide us with an understanding of data protection issues and give an insight into transparency. This work has influenced the design and use of all parts of the toolbox, and has been reported elsewhere (Wilson & Fletcher, 2010).

We already have tools in the WeGov toolbox to enable us to collect publicly-accessible comments and related data from social networks and other web sites; and also to analyse the comments to summarise their subject matter and the opinions expressed in them. A description of one of our analyses can be found in Sizov (2010), which describes the discovery of "latent topics" from a collection of social network postings, which form a summary of the debate, together with the highlighting of key posts and the opinions represented in the posts.

We need to provide a starting point for data collections, and for this we need to determine upon which sites the discussion is taking place. The work described here addresses this, motivated by a use case from external end users, namely to be able to find out where a news article is being discussed. This use case is discussed in detail in the next section, which also sets out the research questions we aimed to answer. Following this, we describe our strategy to answer the research questions. We then describe the results for an example case study showing and how it may be used to benefit by its target audience.

1. Background & Problem Statement

During initial meetings with external end users, a particular need of WeGov's target users, governmental policy makers, was requested. This is the gathering of citizens' opinions as feedback to a particular statement by a politician. The first WeGov prototype covered this scenario as a basic use case. Here, the policy maker posts a statement into a social network, collects the citizens' feedback (where it is publicly available) and runs the analysis components on the feedback. The result is a summary of the key themes and opinions over the sum total of the citizens' comments (Wandhoefer et al, 2010).

The initial toolbox was presented to 29 office employees working for a parliamentarian of the German Bundestag with the aim of gathering feedback for the further development process (WeGov, 2011). During discussions with them, the consensus was that parliamentarians' posts are unlikely to solicit a large amount of feedback, unless the politician is high-profile: "ordinary" parliamentarians' posts typically generate below 100 comments. They confirmed that the requirement to test citizens' reactions to politicians' statements is important, but they need more comments to provide a statistically significant sample of opinions. A modification of the original use case was proposed by the Bundestag employees, where politicians' statements are covered on the internet through news articles, which are in turn disseminated and discussed by citizens. Figure 1 outlines "The Newspaper Story" which capitalises on the effect of "indirect injections" (Joshi et al, 2010) – this means the politician's statement is disseminated by citizens rather than the

politician. For example, a news article is written around the statement, and this is discussed over many different locations by citizens.

In the example in Figure 1, www.bbc.co.uk published a news article with the headline "State multiculturalism has failed, says David Cameron"[1]. Internet news sites provide the opportunity for readers to share and discuss news articles over diverse internet locations, and thus the story may be propagated and discussed in many places on the internet by citizens. This news article was shared 31,309 times on Facebook and 1,922 times on Twitter.

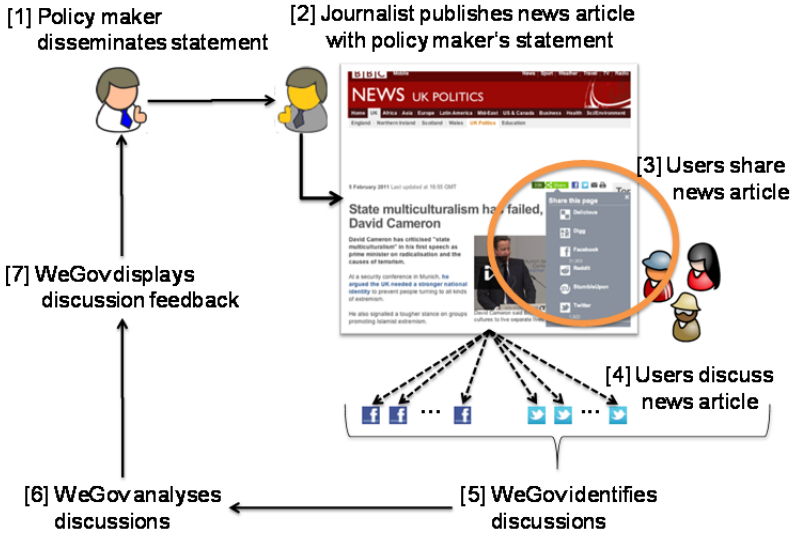


Figure 1: The Newspaper Story

There is thus a vast amount of discussion on this news article, but the challenge is determining *where* it is being discussed. Addressing this challenge (step 5 in Figure 1) is the core of the work described in this paper: to identify the websites where a news article is disseminated, to give the WeGov analysis tools the opportunity to get more user comments as input, which should improve their accuracy. Once we had a good idea of the core problem to be addressed, and began to consider its implications, a number of related challenges presented themselves. These are discussed next, and determine our requirements.

Once people start sharing the news article, discussion spreads out over different sites. This adds a new dimension to the requirements – it would be most helpful to the policy maker to track where the discussion on an article occurs over time from initial publication of the article.

News events are not usually covered with a single news article on one website – it is more likely that there are multiple articles written from different perspectives in different newspapers and on different websites, and each generates their own set of comments from the citizens that read them. In addition, news events develop and multiple articles are written, adding new developments and analysis. This adds a further requirement: to be able to track reactions to multiple articles, and to group them into sets, so they can be presented to analysis in logical groups.

[1]BBC news website. News article "State multiculturalism has failed, says David Cameron", 2011. URL: http://www.bbc.co.uk/news/uk-politics-12371994 (Retrieved 20 Nov 2011).

It is also probable that the policy maker does not know the exact news articles they wish to track, or they only know of a subset of articles, so a related requirement is to enable searching for news articles to "bootstrap" the tracking of reactions to them.

Finally, policy makers are often specialists in, or are responsible for, a certain discipline or topic area, and it would be most helpful to them to determine key sites that are worth monitoring for general discussion and ideas around this topic.

Given the problem statement and the requirements above, a number of research questions arose:

1. How can we find out where a news article is being discussed?
3. How do the discussions' locations change over time?
4. How can we track a news story containing many news articles?
5. How can we find news articles related to a press release or an MP's statement?
6. Which are the important places a policy maker needs to monitor for discussion of items relevant to them?

2. Strategy

This section begins by outlining briefly how we addressed the research questions. After this, details pertaining to specific challenges for each question and their solutions are discussed.

- To address research question 1 (how to find where a news article was being discussed on the internet), we proposed a strategy whereby we perform internet searches for references to the news article and store the results in a database. The assumption underlying this strategy is that if the news story is referenced in a web page (i.e. it is returned as a hit in an internet search for the news article), then that web site has at least some relation to the news story.
- To track how the discussions' locations changed over time (research question 2), we proposed to repeatedly (automatically) execute the same search on a regular basis, and store these results along with the original results in the database.
- Tracking a news story containing multiple articles (research question 3) is a matter of grouping searches and results together into a set for the story. This is simply a question of management of the searches and results; and maintaining links between sets of results, searches, news articles and news stories. We proposed to utilise relational database patterns to maintain these links - databases are well suited for this task.
- If we do not know the URL of the news articles we want to track, we will need to find the articles themselves (research question 4). For this we have proposed that we search selected newspapers' websites for keywords from the press release or MP's statement. The result set should be links to articles about the press release etc. Once we have this set of articles, they can be used as input to the strategies above.
- Finally, in order to determine useful sites for general monitoring (research question 5), we have proposed that we analyse groups of search results, to determine which web domains are most frequently featured, and how they are ranked.
- We also determined that we should be able to select arbitrary sets of search results for this analysis, so that we can determine the best sites given any set of results, from one single set to all sets. This enables maximum flexibility to "data mine" the search results. We should be able to

determine their own groups – for example multiple stories may be related because they are about a similar subject area. Grouping the results from these and analysing them produces a set of web domains pertinent for that subject area.

2.1. Searches

There are two main types of search. The first type is a search for comments and references to news articles. This is the major form of search (addressing research question 1, to find where news articles are being discussed), and returns result sets containing ordered hits for references to the news article. The second search type is a search for news articles given the text or keywords of an MP's statement. This is used when the news articles are not known, and the results can feed into the search for comments and references to news articles. This addresses research question 4.

2.1.1. Searches for Comments and References to News Articles

The basic strategy we chose for this search was to automate data collection based on Google searches for places where the news article occurs, and to store the search results (hits) in a database. The same searches can be repeated periodically over time, and differences in where the news item is discussed can be highlighted.

The first thing we did was determine how to search the internet, and the solution was straightforward: we chose to use Google because it is the most popular search engine in the West. It has by far the largest market share of the search engines[2], and appears to be the de facto choice for most users of the internet. As such, it is very likely to be used by many people who may want to comment on a news article – if someone wants to comment on a news article, they are likely to either:

- comment directly on the news site itself;
- comment on a website / SNS / forum / blog they already know about; or
- Google search for the news article to find places where it is being discussed.

Therefore using Google, we will find pages with comments from people who do any of these actions. We chose to perform two Google searches for each news article: firstly the article's headline and secondly the news article's web page URL. This enabled us to capture references to the news article (when the URL is quoted in a referring text), and the more general search for the headline, which uses natural language processing such as stemming, removal of stop words and fuzzy matching.

Google searching returns references to web pages (hits), ranked by Google's proprietary algorithm. The Google ranking gives us a measure of how useful a hit will be. Google's ranking of a search hit is important to us, as it determines the "popularity" of a site in response to the search, and we wish to find the most likely sites where people will go to make comments. Therefore our goal is in line with Google's, to find the "best" sites for the search. We do not know the exact details of Google's search algorithm as it is proprietary, but we do know it is founded upon citations – links to a web page behave as "votes" to increase its rank. What the algorithm contains

[2] See for example Netmarketshare: http://www.netmarketshare.com/

is not important to us, but that it is used and relied upon by a vast user base, is. Google has a vested interest in returning useful hits to its users, and its market share indicates that it is doing just this.

To perform the actual searching, we used a Google Custom Search Engine (CSE) with automated control to repeatedly execute searches. The frequency of searching is configurable, and our initial configuration is that we search at the same time every other day. We created a relational database to hold the search results. For each hit in a search result set, we record the URL of a hit, its domain and the Google rank, as well as the search information such as the search query and the date of searching. This enables us to perform analyses of how the ranking can change over time as well as aggregate analyses to determine the best sites given arbitrary sets of search results.

An example of search results for a news article, collected over a time period, is shown in Figure 2. The figure illustrates the relationship between the news article and repeated searches for its headline and URL, and also illustrates how we can address research question 2 (tracking the changes of discussion location for the news article over time).

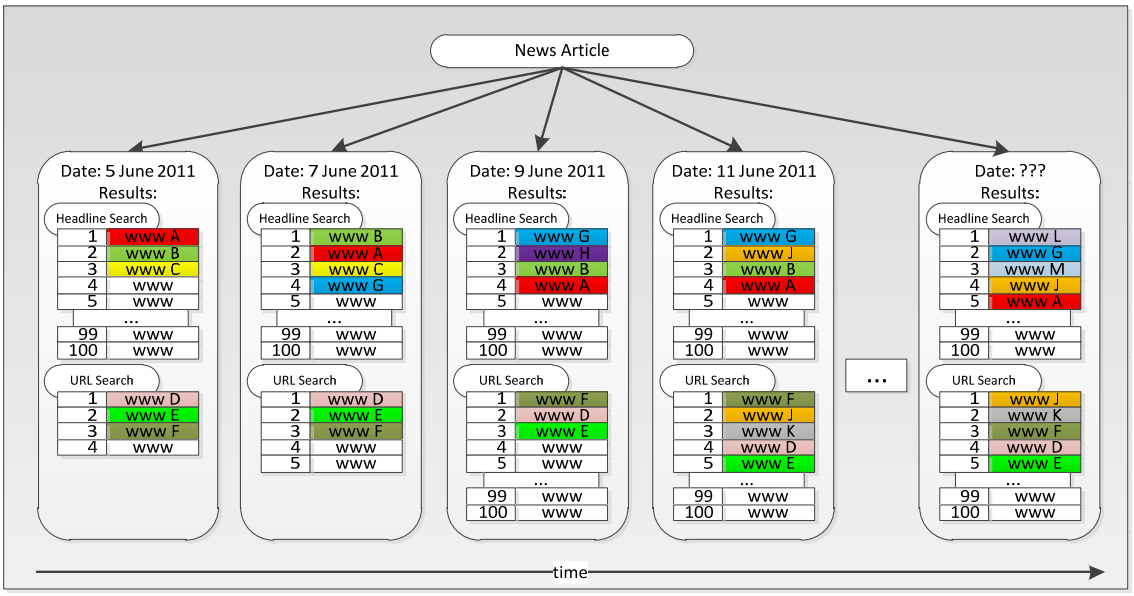


Figure 2: Example News Article, Associated Searches & Results

The sets of hits in the search results above are analogous to the music charts denoting which records are the most popular at a given time. Our "hit parade" is of the top 100 web sites where the news article is discussed. The charts are updated periodically and web sites move up and down the chart according to their popularity. We also get new entrants to the chart and other web sites drop out of the top 100. Figure 2 illustrates this by showing the progress of some example websites –we can see the progress of domain A over time – it starts off at the top, and then drops down the list. Domain G is another example – it does not feature until the second search, but then rises to the top before tailing off. Figure 2 also illustrates that we may not get a full set of hits early on in the search, especially for a fresh news article. This is particularly true for the URL search, as it is highly specific and the search engine cannot use any natural language or fuzzy matching techniques to widen the result set. This is a desirable property for our purposes, as it means we are getting exact matches for the URL, and we can see its propagation.

2.1.2. Searching for News Articles

The second search type, where news articles themselves are found (research question 4), is only required when the news articles are not known, or the user wants to see newspaper reports of a particular policy statement. This search also uses Google, but requires that a limited section of the internet is searched – we only want news websites to be searched here. We created a second Custom Search Engine (named here the "Newspaper CSE"), that only searched a sample of UK news and newspaper sites (for example www.bbc.co.uk/news, www.telegraph.co.uk,etc).

The Newspaper CSE can be searched using keywords from a government press release or an MP's statement, and because it is configured only to search newspaper and news sites, the results will be news articles. These news articles can then be used as inputs to the other search type (references and comments) to see where the articles are discussed on the wider internet.

The choice of news sites is customisable – the Newspaper CSE can be altered at any time, so additional news sites can be included. The Newspaper CSE could also be targeted to a specific purpose, for example a subject area, or geographical location (the Newspaper CSE searches could be local newspapers rather than national ones).

2.2. Data Analysis

By utilising different criteria to group the search results, we can answer research questions 3 (tracking the discussion locations for multiple news articles related to a news story) and 5 (a general aggregated analysis showing useful sites given multiple different data sets).

The grouping process may be thought of as a form of data mining known as an OLAP cube[3]. This allows data to be grouped and analysed along different dimensions. The dimensions we can utilise are: story names / keywords / subject areas, web domains, dates and article titles.

An example OLAP cube for a complete story is shown in Figure 3, alongside an analogous set of searches for news articles. Here we are comparing news articles against dates, and showing the ranked set of domains for each article and date.

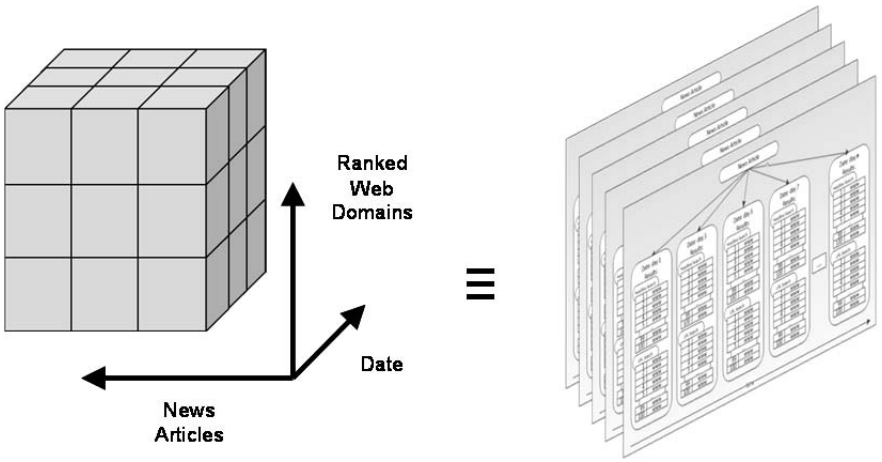


Figure 3: Example OLAP cube

[3] http://en.wikipedia.org/wiki/OLAP_cube (Retrieved 29 November 2011).

We can use slices from the OLAP cube as well – this means we only interested in one value in a particular dimension. For example, if we are only considering one story keyword, we only have a single value on that dimension.

In OLAP-style analysis we often want to aggregate values in a particular axis so we can examine the effect of other axes on the overall result. For example we may wish to investigate the aggregated ranking of all news articles in our story and how it changes over time.

In the right hand side of Figure 3, this is collapsing the five layers into one and finding aggregated rankings in all the hit parades – taking into account all layers. This provides us with an overview of the top sites where the discussions on all our news articles are taking place over time, regardless of the articles.

We thus needed a means of aggregating the rankings, and we attempted different methods of determining the aggregated ranking. We originally attempted simple averaging, e.g. we took all the ranks for a particular web domain and computed their average. The major problem with this method was that it was highly sensitive to the number of records for that domain. If, for example, domain A had a single record at position 1, this would have an average value of 1, because there is only a single record. If domain B had ten records, and nine were at position 1 and the remaining record was at position 2, the average value would be 1.1. Given that the lower the average the better in this example, this means that the consistent high performer, domain B, with 9 top positions, was apparently outperformed by domain A, who only appeared once.

Next, we looked at different weighting algorithms, so as to give more importance to the higher positions, but these suffered similar problems to the straightforward average. It was therefore decided that we needed to take account of the number of occurrences a domain has, as well as its position in each occurrence. What we wanted to find was consistent good performers (e.g. domain B above), rather than ones with few high positions but no other records (who could be considered "lucky" without further evidence).

The Bayesian Average method[4,5]is purpose-built for this task. It reduces the effect of anomalous values by considering the average number of occurrences for each domain as well as the average value per occurrence. It does this by calculating a corrected ranking that takes the number of occurrences a domain has into account using the two following principles: the more occurrences a domain has, the closer its corrected ranking value is to its uncorrected value; and the fewer occurrences a domain has, the closer its corrected ranking is to the average ranking value of all domains. Thus, the more times a domain appears in search results, the more "believable" its scores are.

After the data slicing, aggregation and Bayesian averaging, we have ranked tables of "chart positions" for each domain that take into account the way we have sliced the data, the chart positions and the number of votes for each chart position. Using this aggregation and the OLAP cube technique, we can show the aggregated ranking of multiple search results, for example:

- A single story (e.g. all the searches over time to date for the story)

---

[4]http://en.wikipedia.org/wiki/Bayesian_average (Retrieved 30 November 2011).

[5]http://leedumond.com/blog/the-wisdom-of-crowds-implementing-a-smart-ranking-algorithm/ (Retrieved 30 November 2011).

- The time development of a single story (e.g. where discussions for all news articles in the story change over time)
- Multiple stories (e.g. the user can select related stories to find out where people are talking about them)
- One day (e.g. all searches on one single day independent of story)
- Everything (e.g. all results for all stories to date).

3. Results & Initial Evaluation

We implemented a software tool to perform the searches and analyses described above. We show here an example of its output for a case study of a news article. The main UI of the software tool with the article and its searches displayed is shown in Figure 4.

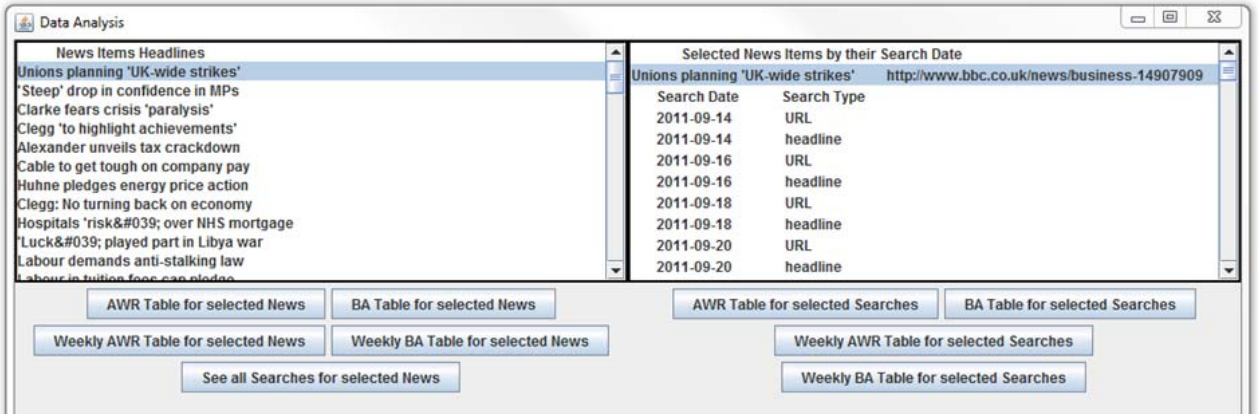


Figure 4: News Story Tracking

The example is based on a single news story, from the BBC news website. This concerned a story about plans for a national public-sector strike, published on 14 September 2011[6]. Figure 5 shows a results page, showing the Bayesian Average ranking, position, and number of occurrences for different domains over all times the news article is searched for (14 September 2011 to 30 November 2011).

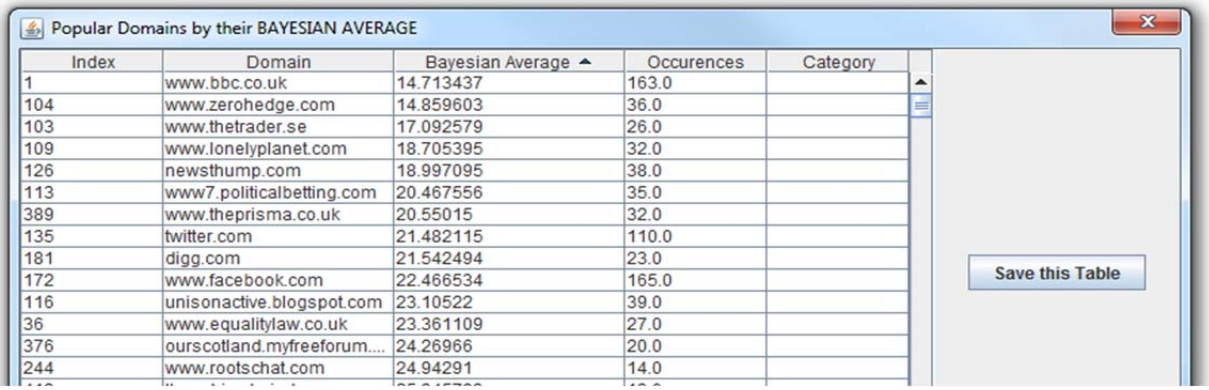


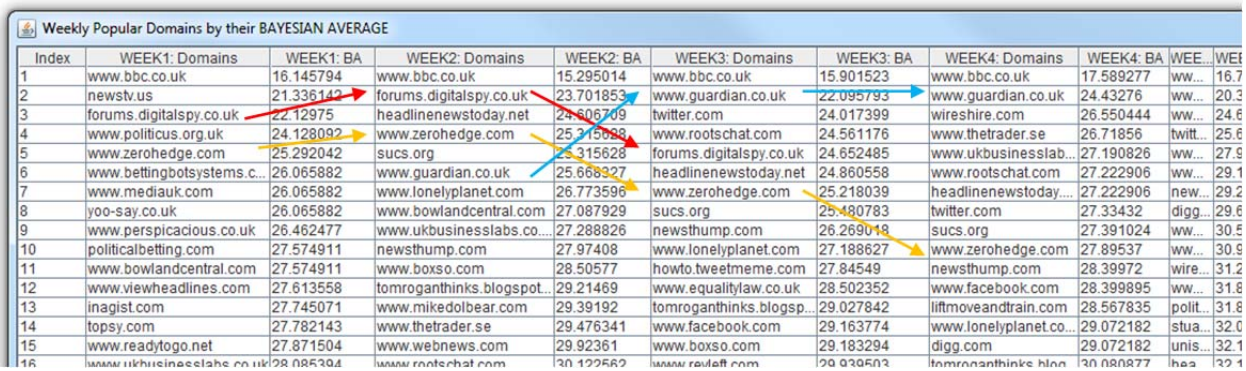
Figure 5: Bayesian Average Overall Rankings

[6]Available at: http://www.bbc.co.uk/news/business-14907909 (Retrieved on 1 December 2011).

Because they reference the news article, the sites can be useful places for the policy maker to watch for opinions on their policy statements. In the results table, the domain is only shown, and we use this as a gateway to the actual results – we click on a domain and we can see the hits associated with that domain. A hit contains the URL of the page, Google's "snippet" and the rank. The table is sorted by the Bayesian Average of the Google rankings. The smallest is first – this means the most interesting sites as determined by Google are at the top. The number of occurrences of each site indicates how many data points were used to compute the overall Bayesian Average ranking, and we can see that there are reasonable numbers of samples (i.e. search results) for each site. We can also see that www.bbc.co.uk is the highest position – this is reasonable and to be expected, since the BBC needs to index its own pages, but is rather obvious. This is no problem– as it is a known good site, so we just look further down the list.

The "category" column allows the user to mark some sites as known good sites, or sites that are not useful. The policy maker can mark a site as "uninteresting" in the "category" column, and it is ignored in this and all other result tables. The data from ignored sites is still collected and analysed, and the blacklist of uninteresting sites can be edited at any time to return the blacklisted sites to the analysis results.

The overall rankings may be broken down by time, and Figure 6 shows rankings per week. The results are grouped per week and are for the single news article. Here we see a more varied set of sites, and their positions change week by week. The arrows in the figure illustrate the movement of some particular domains, and we discuss one of them in more detail next.



Weekly Popular Domains by their BAYESIAN AVERAGE

Index	WEEK1: Domains	WEEK1: BA	WEEK2: Domains	WEEK2: BA	WEEK3: Domains	WEEK3: BA	WEEK4: Domains	WEEK4: BA	WEE...	WEE
1	www.bbc.co.uk	16.145794	www.bbc.co.uk	15.295014	www.bbc.co.uk	15.901523	www.bbc.co.uk	17.589277	ww...	16.7
2	newstv.us	21.336142	forums.digitalspy.co.uk	23.701853	www.guardian.co.uk	22.095793	www.guardian.co.uk	24.43276	ww...	20.3
3	forums.digitalspy.co.uk	22.12975	headlinenewstoday.net	24.606709	twitter.com	24.017399	wireshire.com	26.550444	ww...	24.6
4	www.politicus.org.uk	24.128092	www.zerohedge.com	25.315628	www.rootschat.com	24.561176	www.thetrader.se	26.71856	twitt...	25.6
5	www.zerohedge.com	25.292042	sucs.org	25.315628	forums.digitalspy.co.uk	24.652485	www.ukbusinesslab...	27.190826	ww...	27.9
6	www.bettingbotsystems.c...	26.065882	www.guardian.co.uk	25.668327	headlinenewstoday.net	24.860558	www.rootschat.com	27.222906	ww...	29.1
7	www.mediauk.com	26.065882	www.lonelyplanet.com	26.773596	www.zerohedge.com	25.218039	headlinenewstoday....	27.222906	new...	29.2
8	yoo-say.co.uk	26.065882	www.bowlandcentral.com	27.087929	sucs.org	25.480783	twitter.com	27.33432	digg...	29.6
9	www.perspicacious.co.uk	26.462477	www.ukbusinesslabs.co....	27.288826	newsthump.com	26.269018	sucs.org	27.391024	ww...	30.5
10	politicalbetting.com	27.574911	newsthump.com	27.97408	www.lonelyplanet.com	27.188627	www.zerohedge.com	27.89537	ww...	30.9
11	www.bowlandcentral.com	27.574911	www.boxso.com	28.50577	howto.tweetmeme.com	27.84549	newsthump.com	28.39972	wire...	31.2
12	www.viewheadlines.com	27.613558	tomroganthinks.blogspot...	29.21469	www.equalitylaw.co.uk	28.502352	www.facebook.com	28.399895	ww...	31.8
13	inagist.com	27.745071	www.mikedolbear.com	29.39192	tomroganthinks.blogsp...	29.027842	liftmoveandtrain.com	28.567835	polit...	31.8
14	topsy.com	27.782143	www.thetrader.se	29.476341	www.facebook.com	29.163774	www.lonelyplanet.co...	29.072182	stua...	32.0
15	www.readytogo.net	27.871504	www.webnews.com	29.92361	www.boxso.com	29.183294	digg.com	29.072182	unis...	32.1
16	www.ukbusinesslabs.co.uk	28.085394	www.rootschat.com	30.122562	www.revleft.com	29.939503	tomroganthinks.blog	30.080877	hea	32.1

Figure 6: Search Results - Bayesian Average Rankings per Week

The Digital Spy forums contained two threads that discussed the BBC news article, and was therefore a hit in the search. Digital Spy started off in position 3, peaked in the second week at position 2, and then dropped to position 5 before dropping out of the visible data. The first thread[7] was simply discussing the article, and the first post included a link to the BBC news story. This thread had 120 posts – the first was made on 14 September 2011, and the last was made on 7 October 2011. The bulk of the activity was in the first week. The second thread[8] was a poll about whether people supported the strike, together with opinions given in the thread. The thread

[7]Public Sector Workers Balloted On Strikes. http://forums.digitalspy.co.uk/showthread.php?t=1534213 (Retrieved 8 December 2011).

[8]Do you support the union strikes? http://forums.digitalspy.co.uk/showthread.php?t=1534436 (Retrieved 8 December 2011).

accompanying the poll contained 522 posts, the first was made on 14 September 2011, and the last was made on 19 September 2011. The bulk of the posts were made over 15-18 September.

The pattern is common – an event occurs, and there is a flurry of activity concerning it, which peaks and then tails off. Given that there may be a lag in Google's indexing of the Digital Spy forums, the activity on the forum and the rankings in the table give a reasonable match. The important point is there is genuine and useful debate in this forum. Other forums with debate on this topic were also highlighted by the results table.

This initial evaluation has demonstrated that the approach works, but it is worthwhile assessing the positive and negative aspects of the approach.

- Positive aspects. Firstly, our approach uses Google, the most popular search engine in the western world – its performance is attested by the fact that millions of people use it daily. Secondly, the Bayesian average approach to aggregating search scores has the advantage that it reduces the effect of infrequent anomalous values, resulting in a score supported by the bulk of the data points. Finally, the approach allows the user to see the changes over time in the rankings of the sites discussing the article.
- Negative aspects. The major drawback with our approach is that the ranking of the results are determined by a proprietary and unknown algorithm. However, we can easily adopt another search engine without adjustment of our technique. Another drawback is that in order to see where discussion of a story changes over time, the story has to be tracked from its beginning. Automation can assist here – the user can specify queries to find news stories as they happen, and these can be tracked automatically. A further drawback is that our approach is based on a numerical aggregation of rankings to determine the relevance or popularity of a site pertaining to a particular query – no account is taken of the actual postings on the sites. The lexical analysis of postings is addressed in other aspects of the WeGov project (see for example Sizov, 2010), and this work provides starting points for searches that can provide input to these analyses.

4. Conclusions

This paper has described a method whereby a governmental policy maker can discover where policy statements are discussed, and we have shown some example results validating our approach. Our strategy was to assume that news articles are written about the policy statements, and these are discussed over the internet. To enable us to find these discussions, we automatically scheduled and repeated Google searches for references to news articles' headlines and URLs. We collected the results in a database, enabling us to aggregate and analyse them to produce ranked tables of sites that reference each news article. Using data mining techniques such as the OLAP cube, we can group data so that the result reflects an overall aggregate score, taking into account multiple datasets, averaging out individual differences. We can also examine the differences between datasets, for example how the sites where the article is discussed change over time.

There are two major elements of further work. Firstly, having conducted our initial evaluation, we need to present the work to policy makers, so that they can make comments on the search results, and how the results are presented. Secondly, we need to integrate the software tool with the rest of the WeGov toolkit, so that the results of this work can be fed into more detailed searches and analyses to find out what people are saying.

References

Addis, M., & Taylor, S., & Fletcher, R., & Wilson, C., & Fallon, F., &Alani, H., &Mutschke, P., &Wandhoefer, T. (2010). New ways for policy makers to interact with citizens through open social network sites - a report on initial results. In H. Margetts& S. Gonzalez-Bailon& S. Ward & D. Sutcliffe (Eds.), *Internet, Politics, Policy 2010: an impact assessment*, 16-17 Sep 2010, Oxford, UK.

Hansard Society (2009). *Digital Dialogues Phase Three.* Retrieved 20 November 2011, from http://digitaldialogues.org.uk/reports/digital-dialogues-phase-three/

Joshi, S., &Wandhoefer, T., &Thamm, M., &Mathiak, B., & Van Eeckhaute, C. (2011). Rethinking Governance via Social Networking: The case of direct vs. indirect stakeholder injection. In E. Estevez & M. Janssen (Eds.), *Proceedings of 5th International Conference on Theory and Practice of electronic Governance. New York New York 10121-0701: ACM PRESS, S.* 429.

Sizov, S. (2010) . GeoFolk: latent spatial semantics in web 2.0 social media. In B. Davison & T. Suel& N. Craswell& B. Liu (Eds.), *Proceedings of the third ACM international conference on Web search and data mining, February 04-06, 2010*, (pp. 281-290). ACM.

Wandhoefer, T., &Thamm, M., &Mutschke, P. (2011). Extracting a basic use case to let policy makers interact with citizens on Social Networking Sites: a report on initial results. In P. Parycek& M. Kripp& N. Edelmann (Eds.), *Proceedings of the international conference on e-democracy and open government; 5-6 May 2011, Danube University Krems, Austria, Krems: Ed. Donau-Univ. Krems, S.* 355-358.

WeGov (2011). *Presentation at the German Parliament.*

http://goo.gl/oqgZu.

Wilson, C., & Fletcher, R. (2010). Appendix A, Legal Analysis of Issues Surrounding Social Networking Sites. In W*eGov Deliverable 5.1.*

http://goo.gl/Eu94j

About the Authors

Ruxandra Geana

Ruxandra Geana is a third year University of Southampton undergraduate studying BSc Computer Science.

Steve Taylor

Steve Taylor is a research engineer at the University of Southampton IT Innovation Centre. He is the technical manager of the IST FP7 WeGov project.

Timo Wandhoefer

Timo Wandhoefer is affiliated to GESIS – Leibniz Institute for the Social Sciences. His fields of research contain eParticipation, Social Web and Information Retrieval.