Identifying and characterising public science-related fears from RSS feeds[1]

Mike Thelwall School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478
Rudy Prabowo School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: rudy.prabowo@wlv.ac.uk
Tel: +44 1902 321000 Fax: +44 1902 321478

A feature of modern democracies is public mistrust of scientists and the politicisation of science policy, for example concerning stem-cell research and genetically-modified food. Whilst the extent of this mistrust is debatable, its political influence is tangible. Hence science policy researchers and science policy makers need early warning of issues that resonate with a wide public so that they can make timely and informed decisions. In this paper a semi-automatic method for identifying significant public science-related concerns from a corpus of internet-based RSS (Really Simple Syndication) feeds is described and shown to be an improvement on a previous similar system because of the introduction of feed-based aggregation. In addition, both the RSS corpus and the concept of public science-related fears are deconstructed, revealing hidden complexity. This paper also provides evidence that genetically modified organisms and stem-cell research were the two major policy-relevant science concern issues, although mobile phone radiation and software security also generated significant interest.

Introduction

Public mistrust of science has probably existed as long as science itself. Historical manifestations have varied from popular culture such as Mary Shelley's 'Frankenstein' and H. G. Wells' 'The Island of Doctor Moreau' (Pinsky, 2003; Wilt, 2003; Wolpert, 2005), to mass anti-nuclear power movements (Herring, 2006; Hsu, 2005). In recent years, many scientific issues have become politicised and have given rise to pressure groups, media coverage and public debates. Two of the most prominent have been stem-cell research and genetically modified food, both of which have triggered significant social sciences and ethics research (Hagendijk, 2004; Leydesdorff & Hellsten, 2005; Tait, 2001; Tsai, 2005). The politicisation of science debates has lead to government policy and legislation curtailing researchers' activities in response to public pressure. The importance of this is potentially great (Leydesdorff & Etzkowitz, 2003). Given the key role of research within a profitable modern knowledge-based economy (Etzkowitz & Leydesdorff, 1997; Gibbons et al., 1994), for example in the biotechnology and computing industries, falling behind with a newly-emerging technology is a national potential disaster. Conversely, allowing science to continue 'unchecked' (except for the normal self-regulation process) may cost humanity too high a price if the critics are correct (Chadwick, 2005; London, 2005). In consequence, a large amount of research has been devoted to topics such as the sociology and ethics of individual science policy debates (Klotzko, 2004). It is hence particularly important to be able to identify the next significant science concern debate as early as possible so that research into the social, ethical, legal and policy implications can begin and support politicians to make well-informed, timely decisions. The research reported here is part of an international European Union-funded project

[1] *This is a preprint of an article to be published in the Journal of the American Society for Information Science and Technology © copyright 2006 John Wiley & Sons, Inc.* http://www.interscience.wiley.com/

(www.creen.org) which aims to develop automatic methods to help identify critical public science policy debates.

This paper is devoted to issues surrounding the identification of policy-relevant science concern debates from a type of internet data known as Really Simple Syndication (RSS) feeds. RSS feeds are, in essence, summaries of the updates of web sites or other information sources, such as news feeds. The development and uptake of the technology was driven by blogs and news web sites (see below for a description of how RSS feeds are used). RSS seems set to become a ubiquitous Internet technology: in addition to its original uses, it is also used for podcasting and has become widely adopted by professional web sites, journal publishers and specialist search engines. For example, the PubMed digital repository users can now create customised feeds for specific queries, allowing them to be notified within an hour whenever any new relevant article is added to the database. The triple rationales for using RSS feeds for automatic large-scale analyses are that they allow large collections of blogs and other sources to be efficiently monitored; RSS feeds appear to lend themselves to efficient classification and clustering type tasks (e.g., Wegrzyn-Wolska & Szczepaniak, 2005); and any science concern issue that resonates with the public ought to be reflected to some extent in blogspace.

Industry currently leads the way in the exploitation of blogs, RSS feeds and similar media for large scale analyses. Even the seminal academic work in this area is commercial: it used IBM's WebFountain (www.almaden.ibm.com/webfountain/) for analysing "massive amounts of unstructured and semi-structured text", applying it to a corpus of RSS data (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004). The reasons for IBM's lead over university research were presumably a belief in its potential that was strong enough to finance large-scale research, and the availability of highly developed software (WebFountain). The key commercial application may be in marketing: monitoring large numbers of blogs may allow a company's public image to be continually assessed, as well as the effectiveness of individual advertising campaigns (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004). In addition to IBM, there are also many specialist companies exploiting similar data. One prominent example is the US company Intelliseek, who have coined the phrase Consumer Generated Media (CGM, see www.intelliseek.com/cgm.asp). CGM goes beyond blogs and encompasses any environment where the public can freely post comments, for example newsgroups and email forums. Intelliseek gives away free information about trends in blogspace via blogpulse.com (Glance, Hurst, & Tomokiyo, 2004) and sells the extraction of specific information from their CGM databases to corporate clients. Microsoft is engaging in related research, automatically extracting consumer sentiments from collections of free text (Gamon, Aue, Corston-Oliver, & Ringger, 2005).

In addition to Intelliseek, a number of companies now make available certain types of information gained from large-scale analyses of blogs, or offer free blog search engines (e.g., technorati.com, blogdigger.com, IceRocket.com). The free information includes lists of the top blogs or blog postings (calculated through link counts), lists of 'hot topics', and time series of the frequency with which individual words have been used in blogs. Much of this information could be useful for academic research. Nevertheless, none of the currently available tools allow *context-specific* searches for emerging hot topics or debates. For example, although Blogpulse will report the overall hot topics in blogs and allow users to generate trend lines for their preferred keywords, it will not allow users to search for hot topics within a specific field, such as science debates.

In a previous paper, we introduced the concept of broad issue scanning for the task of identifying and tracking public debates within a broad issue. We used public concerns about science as a case study of a broad issue and introduced and assessed a technique for identifying individual topics from a large corpus of RSS feeds. Essentially, the technique produced a list of relevant words that experienced sudden increases in usage, and used a human evaluator to scan

the list for genuine debates (see below for more details, given as part of an expanded method in the current paper). The results showed that RSS feeds had the potential to be used for broad issue scanning because some debates were identified, but was perhaps too inefficient to be practical without further improvements. The main problem was that only two genuine debates were found from a very time consuming manual examination of 1000 words (Thelwall, Prabowo, & Fairclough, 2006, to appear).

In the current paper we describe what we claim is the first practical system for semi-automatically identifying public science concern debates. It includes several improvements over the previous system that reduces the human effort required to scan the top words, and produces a significantly improved list of top words. In addition we analyse the results of the system in order to be able to theorise about the nature of anxiety-related public science debates in order to speculate about the limits of efficiency for our approach. Finally, we give new information about the composition of the corpus and distribution of activity levels of the feeds.

The first part of this paper gives details of the RSS corpus set up for a project to identify and model public debates about science, together with a discussion of the limitations of quantitative corpus-based approaches. The second part introduces and evaluates our improved word frequency-based time series method for identifying emerging science concerns. The final part of the paper uses stories produced by the automatic method to ground a discussion of science concerns from a taxonomic and social perspective. The questions addressed include: how should 'science' be defined in this context: which types of science concerns are policy-relevant; and how do public concerns about science fit within a wider landscape of science-related fears? This discussion feeds back into a conceptual evaluation of the efficiency limitations of the word frequency techniques used in this paper.

RSS Corpora

RSS feeds are formatted summary information sources accessed via dedicated URLs. There are several different RSS formats but all are XML applications designed to convey a series of chunks of information, called items, within a larger RSS document (Hammersley, 2005; Hammond, Hannay, & Lund, 2004). The items from each feed are continually updated although the feed URL remains the same. For example viewing a news organisation's main RSS feed URL would reveal the top current new stories, but the same URL an hour later would give a different set of stories (items) with new items added and the oldest removed. An RSS reader program (called an aggregator) that checked the URL hourly would be able to notify its owner of each new story, ignoring stories that had previously been seen. RSS aggregators are now in common use and web users can 'subscribe' to a number of feeds related to their interests, relieving them of the need to periodically scan relevant web sites and avoiding the risk of missing information on frequently updated sites. In addition to news feeds, the RSS concept is widely supported by blogs and frequently-updated sites, and seems to set to become an integral component of professionally-designed sites. Alternative innovative applications are also being developed. For example, academic publishers' digital libraries now often offer an RSS feed for each journal, reporting summaries of each new issue as it becomes available; there are now also search engine based RSS feed services (Britt, 2005). The attraction of RSS for users is the time it saves for those who wish to monitor favourite web sites or blogs (Notess, 2002). The attraction for site owners is that RSS feeds are low bandwidth because they are typically terse, graphics-free summaries. Moreover, as RSS readership has expanded, failure to maintain RSS feeds may drive away users (Smith, 2005).

RSS uses a complex XML document structure that would be troublesome for humans to generate but can be built into web authoring software to make RSS feed generation automatic. Blogs are particularly suited to RSS because they are database-driven software applications

that are periodically incrementally updated with new postings. It is a relatively simple matter to extend blog software to generate parallel RSS feeds with no additional user effort.

RSS Corpus Generation

Whilst a small collection of RSS feeds may satisfy the needs of an individual, a large collection may be used to obtain new knowledge from the aggregated information, which is a type of data mining (Han & Kamber, 2000) or text mining (Berry, 2003). For research purposes the method used to create a corpus of feeds is relevant: i.e. the sample selection. It is particularly important because RSS feed sources are heterogeneous, varying from diary-style blogs to international news agencies: a legacy of complex historical origins (Gill, 2005). Unfortunately, however, there is no universal RSS feed register and no other way to generate a complete list of feeds. Hence it is impossible to generate a genuinely random selection and so, in practice, any RSS corpus will be a convenience sample (Kalton, 1983; Sousa, Zauszniewski, & Musil, 2004). The following methods can be used to identify feeds.

- Manual browsing of the web or of sites relevant to the corpus purpose.
- Querying databases of RSS feeds such as CompleteRSS.com.
- Querying commercial search engines, such as with Google `filetype:rss` searches.

All the above were used to generate the corpus reported in this paper. The overall objective was to get as many feeds as possible irrespective of type. The manual browsing was used to identify feeds that might contain science concern information, such as those from environmental pressure groups. A total of 19,587 feeds were found and automatically monitored from February 1, 2005. A previous day of feeds, from January 31, was collected but not used because of the uncertain date of origin of its items: some appeared to be almost a year old in stagnant feeds. The data reported in this paper ran to October 2, 2005.

Figures 1 and 2 summarise the activity of the corpus feeds. This is important because the effective size of a corpus is the number of active feeds. The update frequency of the feeds is also of interest as background corpus information. From Figure 1, the active size of the corpus varied daily with the following overlapping trends.

- A weekly cycle with significantly fewer postings most Sundays.
- A gradual decrease in corpus activity from an average of about 3,500 feeds per day at the start to 2,500 at the end, with a possible levelling out at the end.

The decreasing trend is an issue for long term analysis: it is natural to expect some feeds to die out, particularly the feeds of personal blogs, and so this shows the need for long term analyses to periodically refresh their corpus by adding new feeds. The graph also reveals occasional spikes, mainly caused by technical faults: computer and network crashes causing days of feeds to be lost or not collected, with restarts recording ‘false’ new extra items from previous missed days. These spikes could be removed but this would cause data loss. The variations in the data point to the necessity to generate time series based upon relative word frequencies rather than absolute word frequencies (see below).

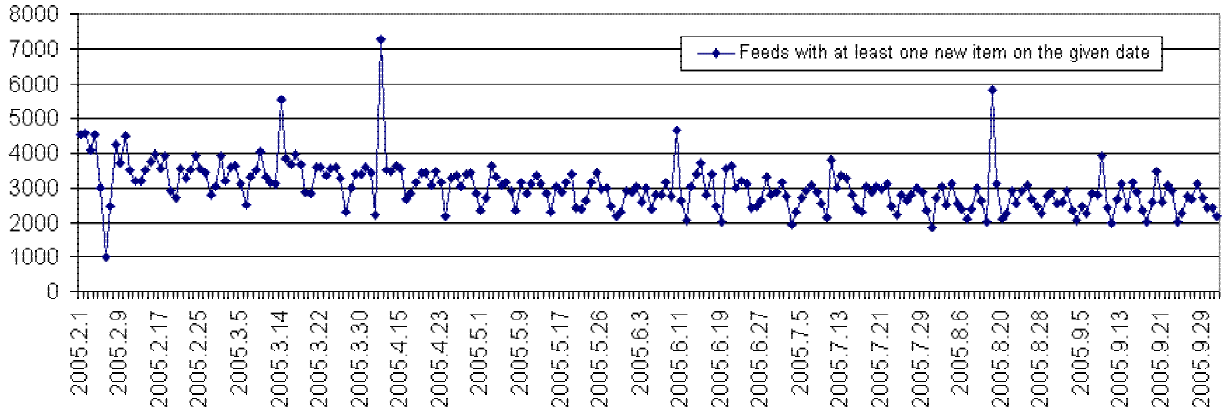

Figure 1. The number of updated feeds per day.

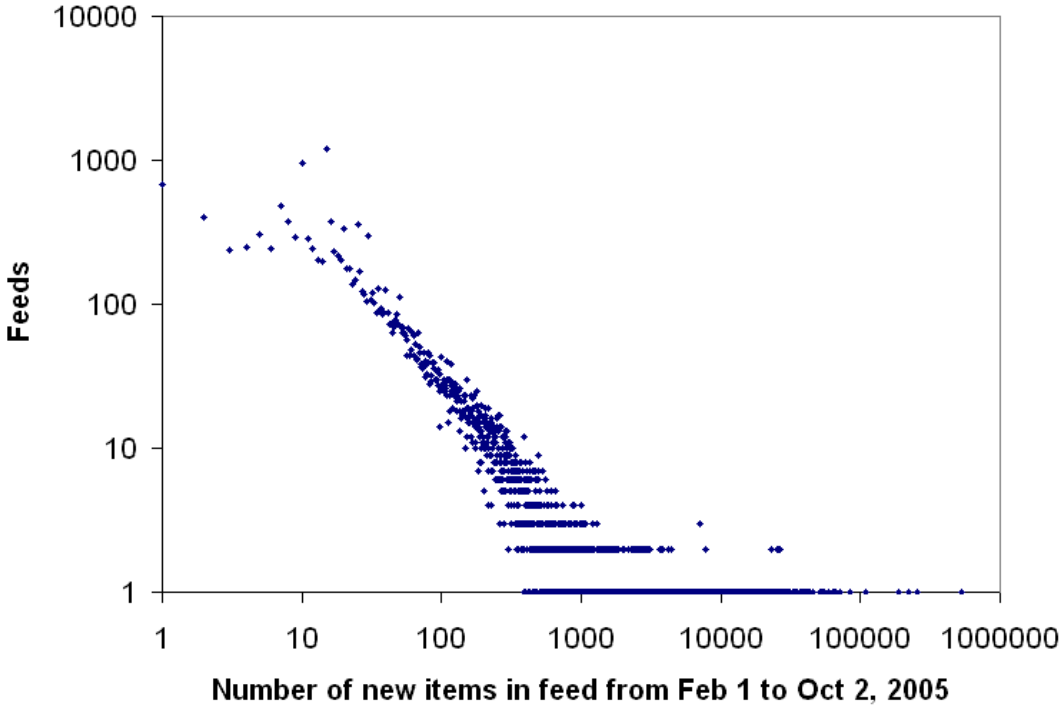

Figure 2. The update frequency of the corpus feeds. An additional 270 feeds posted no new items.

As the graphs show, the corpus contained a wide variety of levels of activity in its feeds, with a few being highly active and many with only occasional posts. Some 50% of the items in the corpus occur in feeds that posted 30 or fewer new items and 50% occurred in more prolific feeds. Importantly for trend detection, then, the corpus was not dominated by prolific feeds. Figure 2 shows the power law distribution that is common in web data (Adamic & Huberman, 2000; Pennock, Flake, Lawrence, Glover, & Giles, 2002; Price & Thelwall, 2005; Rousseau, 1997). This suggests a 'rich-get-richer' effect (Merton, 1968; Zipf, 1949) with a direct feedback mechanism that helps prolific feeds to increase in size. The nature of such a mechanism is not clear for RSS feeds, although it is possibly a side-effect of varying business sizes: successful companies gain more resources and hence can produce more information. Note that the science concern issue identification procedure described below minimises to some extent the impact of highly active feeds. The outliers on the left of Figure 2 are suspicious and suggest the existence of software that periodically generates updates to RSS feeds. For example, the most frequent number of new items per feed was 15, which occurred

for 1,208 feeds, when about 300 would be expected, given the rest of the data. This was found to be because of the setup of the Movable Type blog management software.

Corpus Composition

Within the whole corpus a wide range of sources were present, including just over 6,000 feeds from five major blog sites such as blogstreet.com, suggesting that at least a third of the corpus contained personal blogs. A small random sample of feeds with different activity levels is characterised below to illustrate the types present. Whilst the primary aim of the corpus was to capture public reactions to science concern stories through blogs, the high activity news-related sites are useful to help identify the genesis of stories. This is especially relevant given that almost all significant stories are covered and interpreted by the media, providing a primary information source for many people most of the time (e.g., Bucher, 2002). Nevertheless, there appears to be a growing influence for political blogs like www.instapundit.com, which blur the line between traditional mass media and individual opinions (Gill, 2004; Matheson, 2004), and there are cases where news stories have gestated in blogs before being addressed in the mainstream media (Trammell & Britton, 2005). Of course, many blogs completely ignore news events because their owners are uninterested in current affairs or they have a specific remit for a particular topic (e.g., Bar-Ilan, 2005).

- 529,184 items: http://www.fatwallet.com/rssfeed.php: A site where members can post information about where to find bargains. Its RSS feed contains many very short summaries of member posts.
- 2,889 items: http://www.colorsandpaints.it/rss.xml: The feed of a portal web site for news and links related to paint products.
- 30 items: [URL withheld] A personal US university student blog with general comments about anything he thinks is interesting.
- 15 items: http://www.ofhills.com/blog/index.rdf: A musician's blog advertising his music and commenting on music.
- 8 items: [URL withheld] A personal Taiwanese master's student's blog about his life and travel in Taiwan.
- 1 item: http://www.soulmatesdatingsite.com/dating-sites-by-term/Latin-Lesbian.xml: one page in an apparently database-driven directory of dating sites, which appears to be a search engine optimisation site, i.e. primarily designed for search engine ranking purposes rather than for human readers.

This sample shows to some extent, and the results below show to a greater extent (e.g., Table 1), a strong US influence. This is likely to be related to the development of commercial applications and the 'findability' of US-created RSS feeds (Vaughan & Thelwall, 2004) in addition to heavy US use of blogs. In addition to this inevitable geographic bias, it is known that bloggers are not typical citizens (Lin & Halavais, 2004), for example with students apparently being over-represented in the above list.

Science concern issue identification

Previous research has shown that topic discussions in blogspace are often of a bursty nature (Gruhl et al., 2004; Kumar, Novak, Raghavan, & Tomkins, 2003, 2004). Depending upon the nature of a topic, it may be discussed only once, may be repeatedly "chattered" about with occasional bursts of intensive discussion, or may exhibit slow growth over time. Bursts of blog discussions seem to be almost always triggered by external events such as news stories. Blogspace has many features that promote the discussion of issues. These include the ability to post "comments" on others' blog postings, or to link to an item in another blog to discuss it. Perhaps most significantly, the informal, community-spirited culture of blogs (Bar-Ilan, 2004;

Blood, 2004) can promote an environment in which personal feelings and opinions may be more freely expressed than in other formats that are easily accessible to researchers, such as questionnaires (Fujiki, Nanno, & Okumura, 2005; Huffaker & Calvert, 2005; Viégas, 2005). In contrast to news sources, which may promote mediated or acceptable truths (Seaton, 2005; Herman & Chomsky, 1988) in standardised forms (Entman, 1993), and academic sources which promote academic truths (Fuchs, 1992), blogs contain opinions which frequently lack quality control or wide authority (Cronin, 2005) and hence may collectively give insights into the popular psyche or growing minority grassroots movements (Kim, 2005). Of course, the web can also provide a mechanism for grass-roots mobilisation over political issues (Van Aelst & Walgrave, 2002). The concept of *resonance* is important here (Gruhl et al., 2004). Whilst the typical blog posting never attracts any form of feedback, one that generates a significant degree of activity may be said to resonate with its audience. Identifying the science concerns that resonate in blogspace could give useful early warnings to science policymakers and social studies of science and technology (SST) researchers.

One way to identify resonating blog posts is to use the link structure of blogspace: highly commented posts will have many inlinks from other blogs. Some web sites use links in this way to identify the most popular blogs and blog postings (Glance et al., 2004) and links can also be used for connectivity analyses (Adar, Zhang, Adamic, & Lukose, 2004; Kumar et al., 2004). There are three drawbacks with this method, however, which make it inappropriate for our purposes.

- Link analysis requires downloading and processing full blogs rather than the much smaller RSS feeds, creating a much more significant data collection and processing exercise.
- Links for facilities such as trackback are not universally implemented in blogspace and links are relatively sparse (Gruhl et al., 2004). This is not an issue for high profile stories with many comments and inlinks, but is a problem for others.
- Links will partly measure the influence of a posting and partly that of the poster: widely read blogs naturally generate a higher level of interest, inflating the apparent resonance of their contents.

Text analysis of RSS feeds avoids the first two problems above and probably suffers less from the third than links. The disadvantage of text analysis is the loss of the network structure given by links.

In addition to a small-scale method designed to identify concerns from individual feeds (Fukuhara, 2005), two alternative methods have been implemented to identify science concerns from a large corpus of RSS feeds.

- Natural language processing. Feeds are processed to extract noun phrases and high frequency phrases identified (Prabowo & Thelwall, 2006, to appear)
- Individual words. Feeds are processed to extract words and high frequency words identified (Thelwall et al., 2006, to appear).

Both of the above methods have advantages and disadvantages. Noun phrase identification is extremely time consuming and may incorrectly process new words that emerge during a debate, e.g. frankenwords invented during the genetically modified food debate (Hellsten, 2003; Thelwall, Vann, & Fairclough, 2006, to appear). Word identification may miss significant phrases if individual component words are commonly used in other contexts, as in the case of "stem cell". In this research the more inclusive word frequency approach is used to help a descriptive, intuitive understanding of the data and procedures as well as to help the ultimate goal of fast real-time identification of emerging science concerns.

Science concern term identification method

The science concern issue identification procedure is as follows, where all non-automatic steps are flagged.

1. [semi-automatic step] Generate a large corpus of RSS feeds.
2. Monitor the feeds for an extended period of time, saving all unique items from each feed.
3. For each day and each feed, extract the items containing at least one word, including plurals, from the following list (science, scientist, scientific, research, researcher, researching, researched) and at least one word, including plurals, from the following list (threat, threaten, threatened, fear, afraid, worry, worried, concern, concerned, frightened, scare, scared, risk, risked, risky, danger, dangerous). This sub-corpus is referred to as the science concern corpus.
4. For each word, generate a time series of the production of active science concern *feeds* that contain the word, selecting words that exhibit an 'x day burst', defined as x consecutive days during which the proportion of feeds containing the word is five times higher than the previous average (5 was heuristically determined; see below for the value of x).
5. For each selected word calculate its 'step size', defined as the (minimum) increase in the proportion of feeds containing the word during its three day burst.
6. Rank the selected words in decreasing order of maximum step size during the period.
7. Remove temporal words: e.g. days, months and their abbreviations.
8. Cluster together words that appear to be discussing the same event. This is heuristically determined: for each word, all words of lower frequency are checked and those that co-occur in at least 30% of items on the peak day of the original word are determined to be part of its cluster.
9. For each word, extract a snippet (or key word in context) from the corpus, which is a section of text around the word in an item on its day of peak occurrence. This is similar to the snippets returned in commercial search engine searches.
10. [Manual step] Human classifiers assess the top spike words for the underlying topic that generated them, producing a list of the top science concern stories in the corpus.

The first part of the above procedure (steps 1-6, 10) is the same as previously used (Thelwall et al., 2006, to appear) except that the proportion used in step 4 is calculated per feed instead of per item because individual prolific feeds could otherwise dominate the results. Steps 8 and 9 are designed to give extra information to the human classifier to help them scan long lists of words in a spreadsheet and quickly identify both duplicate words and the reason why each word occurred. Ambiguous cases would still need to be checked by a search of the full corpus (a facility provided in our software). Two other minor changes are an increase in the corpus size with extra dates from April to October 2005, and replacing the words "debate" and "debated" with "threat" and "threatened". Note that plurals were converted to singular in the analysis but no other word stemming was used (e.g., Porter, 1980). Moreover, the burst detection method (step 4) seems reasonable for the evaluation in this paper because it has given some positive results in the past, but future research is required compare it to other standard event/topic detection methods to identify the best one (Smith, 2002; Swan & Allan, 2000; Yang & Pedersen, 1997).

Research Questions

This is still in the early stages of research into extracting information from RSS feeds and hence it is appropriate to have descriptive questions as well as empirical hypothesis testing. We

have four research questions: a specific hypothesis about the methods used; a general hypothesis about the overall efficacy of the approach; a theoretical question about public science debates; and a classification of public fears of science found in the corpus.

1. Does feed-based counting reduce the number of duplicate stories found in the top science concern words? Based upon previous research (Thelwall, Prabowo, & Fairclough, 2006, to appear) this seems likely. Nevertheless, despite all the previous research investigating blogs and RSS feeds, this kind of aggregation has never previously been demonstrated to be effective, nor has it been discussed elsewhere. Moreover, item-based counting appears to be used in the popular blogpulse.com tool. We suspect that commercial RSS and blog analysts probably already apply this method for some analyses but do not report it: nevertheless it is important to establish the superiority of feed-based analysis in an academic context.
2. Is there a significant increase in the number of public fears of science stories represented in the top 200 science concern words, using feed-based counting as compare to item-based counting? Here we do not have a natural quantitative operationalisation of 'significant', e.g. 5% or 10%, although anything less than the previously reported item-based result of 0.2% (Thelwall, Prabowo, & Fairclough, 2006, to appear) would be a failure. This question therefore has to be evaluated qualitatively.
3. What kinds of stories are significant in the science concern corpus according to our method? This is an exploratory, qualitative question. Its purposes are (a) from a social perspective to understand public science debates better; and (b) to build intuition about how our algorithm could be improved (e.g. could we modify it to reliably exclude a class of unwanted public science debates that are not public fears *of* science) and the limitations of word frequency approaches.
4. What are the common current types of public fear of science, as reflected in RSS feeds?

Results 1: Duplicate stories in item-based and feed-based time series

The result of the above science concern term identification method is a ranked list of words in the science concern subcorpus that have experienced a significant increase in use of at least x days' duration, together with suggestions about other similar (clustered) words and key words in context snippets. The new feed-based method was compared with the individual item-based method using both fully automatic and manual methods. For an effective comparison we used x=3, which had previously been determined as optimal for item-based counting. Our experiments suggest that x=1 is optimal for feed-based counting but we used x=3 for both in order to make the case against item-based counting as convincing as possible. The automatic method used the clustering suggestions from the algorithm and regarded all except one element of each cluster as a duplicate. In each case the highest-ranked term in each cluster was regarded as the non-duplicate. Based upon this premise, item-based counting produced 201 non-duplicate terms out of the first 1,000 and item-based counting produced 287, so the new method was clearly superior (the result for optimised feed-based counting, with x=1 was even higher: 343). Figures 3 and 4 are precision-recall graphs for these data.

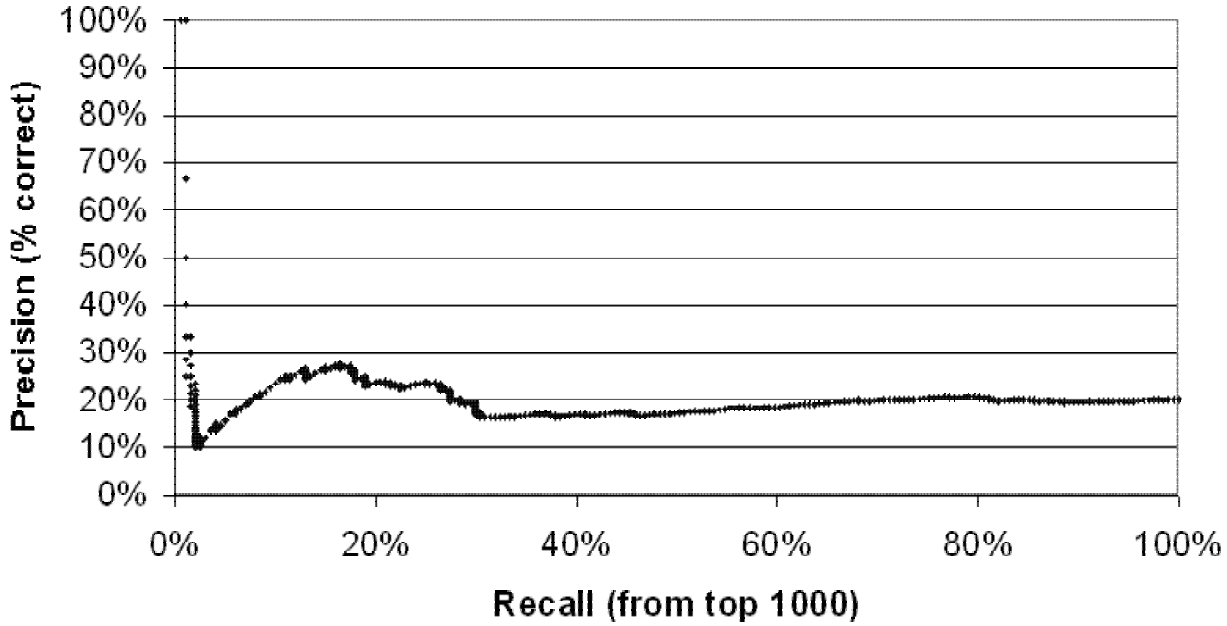

Fig. 3. Precision-recall graph for non-duplicate terms (item-based counting).

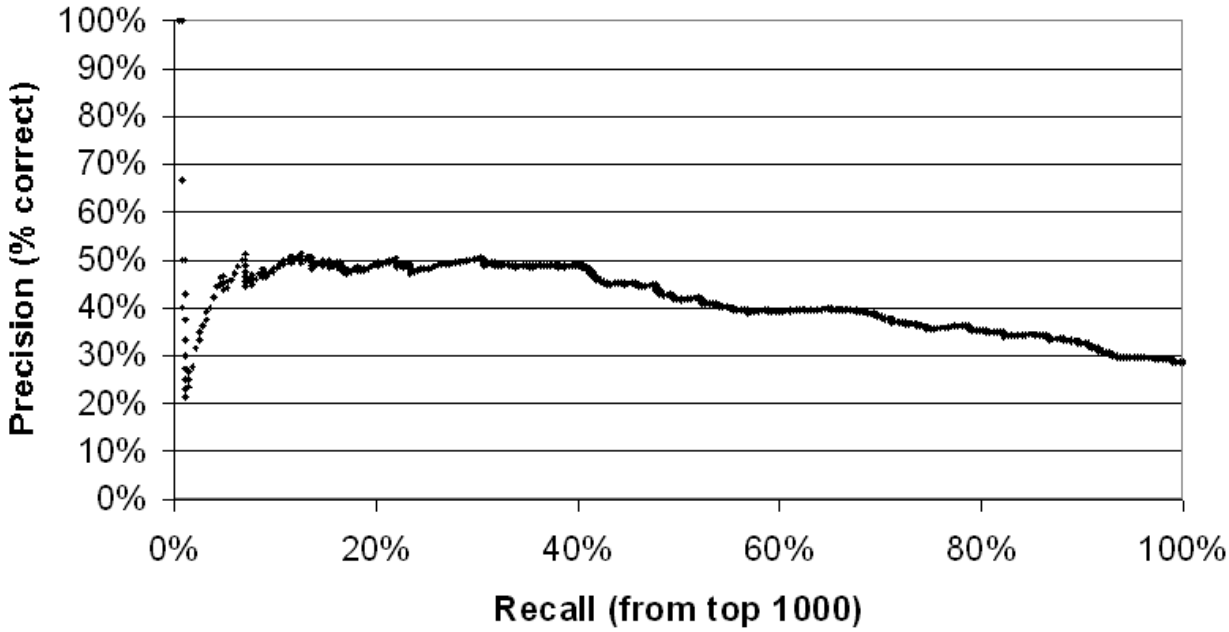

Fig. 4. Precision-recall graph for non-duplicate terms (feed-based counting).

The new feed-based method was also manually compared with the item-based method by selecting the top 200 words produced by each method and both authors independently identifying the issue that caused the word to appear (if a single cause was found) by reading and comparing all the items containing the word during the burst period (see below for a more complete description of the classification process). Although the classifiers disagreed significantly in their results, the conclusion was the same in both cases: more duplicates for item-based counting (77.5% and 61.5%) than feed-based counting (44% and 39%). In fact this hypothesis could be confirmed by a cursory examination of the data: the old method was dominated by repeat postings from a few feeds. In conclusion, feed-based counting reduces the number of duplicates in the list of top stories by reducing the impact of bloggers that repeatedly repost similar modified stories.

Results 2 and 3: Identifying public fears of science and other science concern stories

The words produced by the science concern issue identification process described above were associated with clearly identifiable discussions in some cases but in other cases words were in the list by chance. Although it seems unlikely that a word would enjoy a sudden three-day burst of use without a single underlying cause, this occurred frequently because of the huge number of possible coincidences in a corpus of millions of words.

The classification process

The science concern stories produced by the top 200 words were examined by the first author and manually clustered to produce sets of stories that had a similar orientation on the concept of science concern. The clustering did not rely upon the snippets (step 9 of the algorithm) but referred back to the raw data in case the snippets were misleading.

The only predefined cluster type in the classification was determined by the second research question: public fears *of* science. The classification scheme is described below; and was given to the second author to independently classify the same data set (results in Figure 5). In this section the word 'story' is used in place of the more standard 'topic' or 'event' (as in Topic Detection and Tracking (TDT) research: Wayne, 1998) because it is more intuitive for describing the classification outcomes. The word 'story' also has an alternative meaning as the contents of a single RSS feed item, which would be consistent with TDT terminology.

Fear of science, technology or progress A story that discusses a situation in which scientific or technological 'progress' or theory is seen as threatening human life or enjoyment of life. In this context, "science" is defined as the product of research by academics, including social scientists, computer scientists and humanities researchers, or of researchers/developers in industry (e.g. new car, new breed of cattle).

Information about events or the organisation of research A story in which scientists are used to comment on an event that is not directly caused by technological progress, or scientific theories themselves are described or the organisation of science or of scientific companies. E.g. medical stories where there is no implication of a cure; scientific descriptions or explanations of disasters.

Progress in science or technology, from a positive or neutral perspective A story in which science/technology progress was described, even if on a commercial basis, and where the context was neutral or positive. E.g., medical cures, medical diagnoses, research-based health advice, new software releases.

Threat detection A story in which a prediction was made of a future threat to the planet or a substantial sector of society, but not to individuals, with or without the implication that by predicting the threat it could be avoided. E.g. earthquake/flood/economic prediction, but not medical diagnosis technology.

Other None of the above, for example invocations of science to justify political actions.

Duplicate A story previously coded, i.e. associated with a word with a higher burst size. Stories were considered different if there was a clear event that separated them, one which had been discussed in the second but not the first.

Random The word occurs in at least three different stories; and none of the stories account for at least a third of the postings.

The clustering used the concept of 'story' in an intuitive way from the perspective of the first author, with stories typically relating to a single real-world event. We were interested in individual news stories/events rather than broad topics because the goal of the system is to enable daily monitoring of science-related discussions and so the concept of topic would be too broad. Whilst in most cases the decision seemed clear, in others there were alternative reasonable classifications, with decisions likely to vary with classifier perspective. For example

many items discussed aspects of the 2005 New Orleans floods. From a UK perspective this could be seen as one major event with its aftermath lasting for months. Someone in Louisiana may have a completely different perspective, perceiving a series of connected separate events such as: the hurricane warning; the hurricane; the realisation of the extent of the damage; the criticism of George Bush; discussion of alternative schemes for repairing the levees. In word frequency terms different sub-stories may be identified by related words peaking later; for example 'Katrina' before 'Orleans' before 'levee', although 'Katrina' and 'Orleans' would have a high frequency in 'levee' stories, indicating a connection between the events. Occurrence patterns of word frequencies are thus logical choices as heuristics to identify separate stories within an encompassing longer-term issue. Within longer term debates such as stem cell research, the pattern is probably for occasional bursts of discussion around specific related events: "spiky chatter". Similar conceptual issues are faced in TDT research when differentiating between 'events' and 'topics' (www.nist.gov/speech/tests/tdt/).

Figure 5 shows the results of the classification exercise, comparing the two methods used and giving minimum and maximum values for the two classifiers. As an aside, both classifiers found that the single snippets (step 8) were insufficient for a reliable classification and that all relevant items had to be checked and so in future our system will report all relevant snippets instead of just one per term. The feed-based counting method produced significantly more random events that did not relate to single stories. This was due to the smaller numbers involved when feeds were aggregated; a larger corpus of feeds could significantly reduce this phenomenon. In addition, the feed-based method identified more of all types of story except threat detection. The number of successfully identified fear of science stories with item-based counting was due to a few prolific relevant bloggers so the success rate for story identification is highly dependant upon these individuals. This suggests that a simpler way to identify science concern stories would be to manually scan the blogs of a few active concerned individuals, which policy makers may already do. Nevertheless, the graph suggests that this method would miss almost half of the stories.

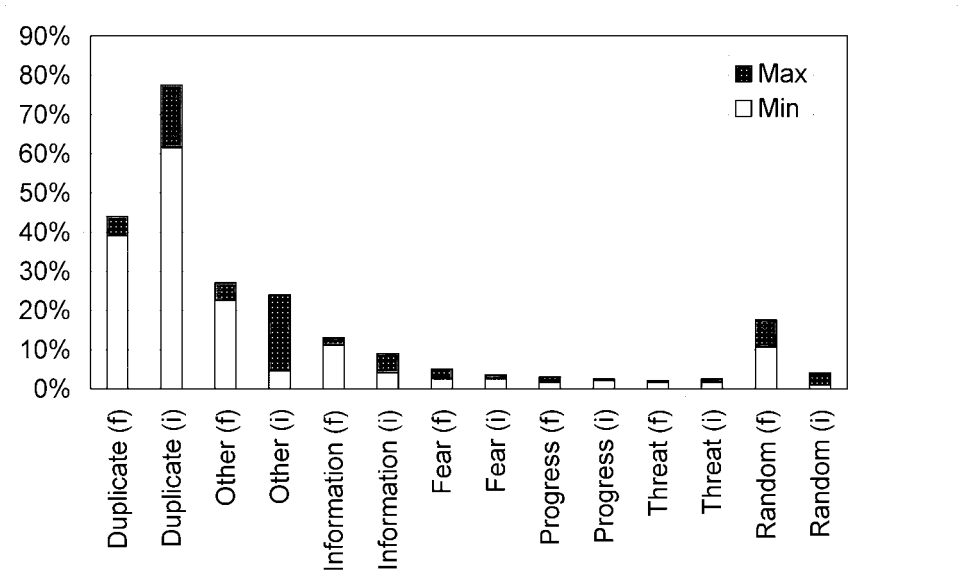

Fig. 5. Classified story types associated with the top 200 science fear words using feed-based (f) and item-based (i) counting.

The large degree of disagreement between classifiers seemed to reflect the fact that some topics could fit into several categories, and that subjective judgements, background knowledge and investigations were required to make a decision. For example, a story about new scientific findings in stem-cell research could have been regarded as a progress story or, since stem-cell research is highly controversial, it could be reported only because of the public fear of this type of research. The concept of science and progress also caused some disagreements. For example, new software could be seen as scientific progress (even if produced outside universities) or as a commercial venture or a routine development (the 'Other' category).

The classification produced a range of different story types that were science-related and concern-related but not public concerns about science. Given the attention in the social sciences for fears of science and criticisms of the scientific method it is interesting to see science frequently portrayed in a positive light or in the authoritative context of providing explanations. In this context, since the stories were identified through explicitly searching for fears or concerns, it is reasonable to suppose that discussions of science in general would tend to be more neutral and more information-related than suggested by the figure. To assess this, a similar exercise was conducted for a science corpus (i.e. just the first half of step 3 above). The results were dominated by the same stories, with some additional information and progress stories. The two top stories that were not captured by the science concern exercise were a NASA probe deliberately crashing into a comet and the discovery of a chimpanzee fossil that might shed new light on human evolution. In addition, hurricane Katrina-related events were more prominent in the list.

In terms of the second research question, the hit rate for public fears of science was 3-6%, an order of magnitude improvement over the previously reported 0.2% (Thelwall, Prabowo, & Fairclough, 2006, to appear), although less than double the equivalent results for the current dataset (Figure 3), for the reasons discussed above. The parameters in the system were optimised for item-based counting (i.e. 3-day bursts) to make the argument in favour of feed-based counting as convincing as possible and the feed-based results could be improved by switching to 1-day bursts.

In terms of the third research question, this is answered by Figure 3. Our objective is to improve the algorithm producing the list of words so that as many of the top words as possible represent public fear of science stories. In some cases the stories were not really about science but used scientific commentators to discuss small aspects of the story (e.g. the New Orleans floods). These could perhaps be eliminated through the use of techniques to identify topics with a low proportion of science-related posts. This would be possible in theory, but would take us away from exclusive processing of the science-fear sub-corpus, and would hence significantly increase processing time to deal with the larger corpus, perhaps even the full corpus. It is also clear that there are some story types that are not public fears of science, but which nevertheless closely involve fear and science. It seems unlikely that our pure word frequency algorithm could be improved sufficiently to exclude these from the list of potential science fear stories. Nevertheless, a more sophisticated natural language processing (NLP) approach might be more successful, although this would also significantly increase processing time because NLP algorithms are relatively slow compared to word frequency approaches. In summary, more efficient methods of producing lists of public fears of science seem available, but mainly at the expense of significant increases in computing time.

The main limitation of the assessment of our method is that we have only assessed precision (the percentage of 'correct' or desired results out of the 200 terms) and not recall (the percentage of stories found out of all genuine public fears of science debates) because there is no authoritative database of public fear of science debates. Hence we do not know how many debates the method will have missed. It is certainly likely to miss some, for example those

described with words that are already very common in the corpus. Hence the method cannot claim to be comprehensive. A second limitation is that we have made no claim that any of the stories identified are policy-relevant. In fact the ultimate test will be how early the system could detect the next major policy-relevant public fear of science debate, but these are quite rare, with perhaps less than one per year and so we have had to develop the system based upon testing on the wider class of all public fear of science debates. A third limitation is that we only used two classifiers, and there was a relatively low degree of inter-classifier agreement. We acknowledge that the classification of this kind of data is highly subjective. This classification hence primarily represents the first author's perspective on the data rather than a tightly defined operationalisation of the classes used.

Results 4: Public fears of science

Table 1 reports some of the stories found that were potential public fears of science, even though not all were classified by us in this way. In order to understand the stories classified as public fears of science, the classification process needs some deconstructing. The term 'science' has many different meanings. In the classification scheme it is interpreted in the wide sense of an activity that might be described as science, technology or innovation, whether occurring in universities, industry or elsewhere. In modern knowledge-based economies it is increasingly difficult to separate scientific and non-scientific activity. Moreover, researchers that do not employ the scientific method may also be called scientists in a broad interpretation of the term. These may also cause public concern, as in the case of an economist with influential theories that could be a threat to future prosperity (see Table 1). Note also that classification can be a political act (Bowker & Starr, 1999): one story was classified as primarily political even though one side of the debate claimed that it was about science.

Some of the stories identified were public concerns about science in the wide sense used but were nevertheless unlikely to be policy-relevant (see Table 1). An example is the story concerning potential bugs in computer network routers. Whilst routers are high tech electronic devices and their failure may cause widespread economic damage it seems unlikely that this will become policy relevant. Essentially, this is because routers are an accepted technology and any failures of new devices could be dealt with through normal legal and commercial processes; this is thus perhaps a 'normal' risk in the context of a risk-driven society (Giddens, 1990). The argument is about whether it works rather than whether the type of technology should be developed and researched, in contrast to the stem cell and genetically modified crops debates. Intuitively, a science policy issue is one where there is a new type of activity or product developed, or a significant change to an existing activity or product is proposed or made. A more minor change to an otherwise existing form of knowledge or knowledge product seems more likely to be dealt with by legislation, although policy changes can be almost impossible to predict, even for experts (Baumgartner, 2006, to appear). From the perspective of improving our algorithm, it seems unlikely that policy relevance could be automatically and reliably diagnosed in public science concerns and so human judgement will always be required to identify genuinely important issues.

Table 1. A selection of manually identified fear of science stories.

Term*	Topic	Comments
rove	Financial	Sudden burst of discussion over fears of the economic theories of Karl Rove, an influential advisor to George Bush, apparently triggered by a New York Times opinion piece.
unlock	GM food	GM vines worry, with a common quote: “French scientist Jean Masson carefully unlocks the gate of a heavily protected open-air enclosure”.
weed	GM food	GM industry puts human gene into rice, allowing it to resist herbicides that can then be used to kill weeds.
connecting	Mobile phone radiation	Research showing that mobile phone use for ten years does not increase risks of developing a tumour in the nerve connecting the ear to the brain.
tumor	Mobile phone radiation	Research showing difference between rural and urban brain tumour risks for regular mobile phone users.
pin	Security of new technology	Research showing that consumers’ pin numbers could be revealed by poor printing.
amid	Software security	Concern over spyware features in a software vendor’s products, with the company releasing a denial “amid” reports of a takeover bid.
cisco	Software security	Former Cisco security researcher claims that there are “fatal flaws” in its routers.
buffer-overflow	Software security	Researcher agrees to no longer discuss claimed Cisco flaws.
disclosure	Software security	Oracle’s chief security officer defends the commercial security record claiming that “the threat of public disclosure” is not needed in order to take action.
frist	Stem cell research	US Senate Majority Leader Bill Frist announces his support for legislation to increase funding for human embryonic stem cell research
academie	Stem cell research	US National Academies' guidelines for Stem Cell research published.
loosen	Stem cell research	Debate on legislation to loosen restrictions on embryonic stem cell research funding
stem	Stem cell research	George Bush threatens to veto stem cell research bill.
korea	Stem cell research	Announcements about human cloning research in South Korea

* The (de-pluralised) word with the highest increase in relative frequency that was associated with the story.

Summary and future research

In this paper we have shown that feed-based counting is superior to item based counting for the identification of public science debates. The feed-based method produced less than half as many duplicate stories than the item-based method. It also produced more public fear of science stories (and an order of magnitude more than previously reported), the ultimate objective of the system. We believe that future researchers using RSS data, or even blog data, to extract time series should use feed/blog based counting rather than item/post based counting. Although further research is needed to confirm that this is appropriate for applications other than science fear story identification, it seems likely that similar findings would apply to most or all applications.

Our method is not fully automatic and uses a human filtering stage. The classification of results suggests that this is a limitation of the system that can be avoided by further improvements in methods, although these would need a substantial increase in computing resources to achieve, either because more sophisticated algorithms would be needed or because

more of the data would need to be processed. In addition, the method would probably never be perfect because of the number of stories that involve public fears and science but are not public fears of science. Moreover, the need to diagnose policy-relevance in debates means that a human decision maker would ultimately be required to scan lists of stories and decide which ones required action.

The corpus used was not a systematic sample in a sense that would allow statistical inference to be made about all RSS feeds, which might be relevant for researchers wishing to generate robust results, perhaps about specific areas of the Internet. A more restricted selection mechanism would be necessary for statistical inference; for example, if only Google-indexed feeds were used then statistical conjectures could be made about 'all Google feeds' based upon the corpus. However, the purpose of the corpus reported in this paper was not to make inferences about blogspace or RSS feeds in general, but to identify public science concerns. For this purpose, the ad-hoc corpus creation method was adequate.

A corollary from the investigation of feed types is that the corpus may be improved by removing clearly irrelevant feeds, particularly those with a high activity level. For example feeds that are primarily for commercial advertising purposes could be removed because these have little value for science policy identification so it is of ethical and practical value to remove them. This may also reduce the occurrence of random causes of 'bursty' words.

An interesting side-effect of the system is the ranked list of debates produced. Although providing no surprises, it is (albeit weak) quantitative confirmation that stem cell research and genetically modified crops are major science concern issues of the day (e.g., Hellsten & Leydesdorff, 2005), together with a reminder of the importance of fully functioning complex computing infrastructures to modern society. Moreover, it sets a wider context in which science also addresses public fears, particularly with regard to medical progress and health advice, and helps us to understand events in the world by providing an apparently expert commentary.

In summary, the ultimate goal of RSS-based science-concern issue identification is to produce a list of stories in which aspects of science or concern are present, with a human evaluator identifying the policy-relevant stories from the list, a kind of computer-assisted environmental scanning (Wei & Lee, 2004). Although our investigation also showed that most relevant stories would be covered by a few active bloggers and so an alternative science concern issue identification method would be to manually scan a few particularly relevant blogs, this may miss unexpected debates that are not high profile when they are first discussed but nevertheless resonate with the public.

Given the results reported here, could RSS feed scanning be of practical use in science policy identification with the tools that we have developed? One of our findings was that many fear of science stories could be identified from just a few topical bloggers, but a policy of checking just these might miss unexpected debates that policy makers need to be particularly aware of. We estimate that it would take about half an hour per day for a daily check of the top 200 words for new science policy debates, and a corpus of about 60,000 feeds could reasonably be collected on a single PC, with a second PC needed for data processing (our current mode of operation). Periodic maintenance would also be needed to update the corpus. The time taken by the computer to produce the results would be about six hours, and so in practice, the system could give its results in the late afternoon based upon data collected the day before. We have designed our system with a dynamic index specifically to speed the process of generating daily updates. Similarly, the system can easily report the top increases for a single target day. Hence, if the sponsors of this research wished, they could obtain a daily report on the previous day's main public fears of science debates for a moderate outlay. Given the importance of new technology to the modern economy and the potentially catastrophic risks associated with failure to respond quickly to emerging issues (e.g., stem cell research, genetically modified

crops) we believe that the technique is both practical and a reasonable investment for a government science policy unit. Many companies have competitive intelligence/environmental scanning (Wei & Lee, 2004) specialists and it is a logical extension for a government to employ similar techniques. Probably, however, the daily reports should instead be commissioned from one of the private blog analysis companies that already serve industry, either using the techniques described above or equivalent proprietary techniques developed by the company. This could be cheaper and provide access to a larger corpus of data. In this case, our results would serve to persuade policy makers of the viability of the general approach. An alternative application, and one that we intend to pursue, is for researchers to combine the use of the new technique to identify science fear debates with other techniques designed to track these debates and to model and understand their propagation dynamics.

Acknowledgement

The work was supported by a European Union grant for activity code NEST-2003-Path-1. It is part of the CREEN project (Critical Events in Evolving Networks, contract 012684).

References

Adamic, L. A., & Huberman, B. A. (2000). Power-law distribution of the World Wide Web. Science, 287(5461), 2115.

Adar, E., Zhang, L., Adamic, L., & Lukose, R. (2004). Implicit structure and the dynamics of blogspace. *Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference*, http://www.sims.berkeley.edu/~dmb/blogging.html.

Bar-Ilan, J. (2004). An outsider's view on "topic-oriented" Blogging. *World Wide Web Conference*, http://www.www2004.org/proceedings/docs/2002p2028.pdf.

Bar-Ilan, J. (2005). Information hub blogs. *Journal of Information Science, 31*(4), 297-307.

Baumgartner, F. R. (2006, to appear). Punctuated equilibrium theory and environmental policy. In Punctuated Equilibrium Models and Environmental Policy (Robert Repetto, ed.). New Haven: Yale University Press, forthcoming 2006.

Berry, M. (2003). *Survey of text mining: Clustering, classification, and retrieval*. New York: Springer.

Blood, R. (2004). How blogging software reshapes the online community. *Communications of the ACM, 47*(12), 53-55.

Bowker, G., & Starr, K. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: The MIT Press.

Britt, P. J. (2005). RSS investors - There's gold in supporting them thar standards. *EContent, 28*(9), 8.

Bucher, H. J. (2002). The internet in crisis: Communication in the case of September 11th. *First Monday, 7*(4), Available: http://www.firstmonday.org/issues/issue7_4/bucher/, accessed November 21, 2005.

Chadwick, R. (2005). Professional ethics and the 'good' of science. *Interdisciplinary Science Reviews, 30*(3), 247-256.

Cronin, B. (2005). Vox populi: Civility in the blogosphere. *International Journal of Information Management, 25*(6), 483-586.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. Journal of Communication, 43(4), 51-58.

Etzkowitz, H., & Leydesdorff, L. (1997). *Universities and the global knowledge economy: A triple helix of university-industry-government relations*. London, UK: Cassels Academic.

Fuchs, S. (1992). *The professional quest for truth: A social theory of science and knowledge*. Albany, NY: SUNY Press.

Fujiki, T., Nanno, T., & Okumura, M. (2005, May 10th). *Differences between Blogs and Web Diaries, Toshiaki Fujiki, Tokyo Institute of Technology.* Paper presented at the WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan.

Fukuhara, T. (2005, May 10th). *Analyzing concerns of people using Weblog articles and real world temporal data, Tomohiro Fukuhara.* Paper presented at the WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan.

Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E. (2005). Pulse: Mining customer opinions from free text (IDA 2005). Lecture Notes in Computer Science, 3646, 121-132.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The New Production of Knowledge*. London, UK: Sage.

Giddens, A. (1990). *The consequences of modernity*. Stanford, CA: Stanford University Press.
Gill, K. E. (2004). *How can we measure the influence of the blogosphere?* Paper presented at the WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.
Gill, K. E. (2005, May 10th). *Blogging, RSS and the information landscape: A look at online news.* Paper presented at the WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan.
Glance, N. S., Hurst, M., & Tomokiyo, T. (2004). *BlogPulse: Automated trend discovery for weblogs.* Paper presented at the WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.
Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). *Information diffusion through Blogspace.* Paper presented at the WWW2004, New York, http://www.www2004.org/proceedings/docs/1p491.pdf.
Hagendijk, R. (2004). Framing GM food: Public participation and liberal democracy. *EASST Review, 23*(1), 3-7.
Hammersley, B. (2005). *Developing feeds with RSS and Atom*. Sebastopol, CA: O'Reilly.
Hammond, T., Hannay, T., & Lund, B. (2004). The role of RSS in science publishing: Syndication and annotation on the web. *Dlib, 12*, http://www.dlib.org/dlib/december04/hammond/12hammond.html.
Han, J., & Kamber, K. (2000). *Data mining: Concepts and techniques*. New York: Morgan Kaufmann Publishers.
Hellsten, I. (2003). Focus on metaphors: The case of "Frankenfood" on the web. *Journal of Computer Mediated Communication, 8*(4), http://www.ascusc.org/jcmc/vol8/issue4/hellsten.html.
Hellsten, I., & Leydesdorff, L. (2005). Measuring the meaning of words in contexts: An automated analysis of controversies about 'Monarch butterflies,' 'Frankenfoods,' and 'stem cells.' *in preparation*, http://users.fmg.uva.nl/lleydesdorff/meaning/measuring%20meaning.pdf.
Herman, E. S., & Chomsky, N. (1988). *Manufacturing consent: The political economy of the mass media*. New York: Pantheon Books.
Herring, H. (2006). *From energy dreams to nuclear nightmares: Lessons from the anti-nuclear power movement in the 1970s*.Charlbury, UK: Jon Carpenter Publishing.
Hsu, S. H. (2005). Advocacy coalitions and policy change on nuclear power utilization in Taiwan. *Social Science Journal, 42*(2), 215-229.
Huffaker, D. A., & Calvert, S. L. (2005). Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication, 10*(2), http://jcmc.indiana.edu/vol10/issue12/huffaker.html.
Kalton, G. (1983). *Introduction to survey sampling*. Beverly Hills: Sage Publications.
Kim, J. H. (2005). *Blog as an oppositional medium? A semantic network analysis on the Iraq war blogs.* Paper presented at the Internet Research 6.0: Internet Generations, Chicago.
Klotzko, A. J. (2004). *A clone of your own? The science and ethics of cloning*. Oxford: Oxford University Press.
Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003). *On the bursty evolution of blogspace.* Paper presented at the WWW2003, Budapest, Hungary, http://www2003.org/cdrom/papers/refereed/p477/p477-kumar/p477-kumar.htm.
Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2004). Structure and evolution of blogspace. *Communications of the ACM, 47*(12), 35-39.
Leydesdorff, L., & Etzkowitz, H. (2003). Can “The Public” be considered as a fourth helix in University-Industry-Government relations? Report of the fourth triple helix conference. *Science and Public Policy, 30*(1), 55-61.
Leydesdorff, L., & Hellsten, I. (2005). Metaphors and diaphors in science communication: Mapping the case of ‘stem-cell research’. *Science Communication, 27*(1), 64-99.
Lin, J., & Halavais, A. (2004, May 18th). *Mapping the blogosphere in America.* Paper presented at the WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, New York.
London, A. J. (2005). Undue inducements and reasonable risks: Will the dismal science lead to dismal research ethics? *American Journal Of Bioethics, 55*(5), 29-32.
Matheson, D. (2004). Weblogs and the epistemology of the news: Some trends in online journalism. *New Media & Society, 6*(4), 443-468.
Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56-63.
Notess, G. R. (2002). RSS, aggregators, and reading the blog fantastic. *Online, 26*(6), 52-54.
Pennock, D., Flake, G. W., Lawrence, S., Glover, E. J., & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences, 99*, 5207-5211.
Pinsky, M. (2003). *Future present: Ethics and/as science fiction*. London: Associated University Presses.
Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130-137.
Prabowo, R., & Thelwall, M. (2006, to appear). A comparison of feature selection methods for an evolving RSS feed corpus, *Information Processing & Management*.
Price, E., & Thelwall, M. (2005). The clustering power of low frequency words in academic webs. *Journal of the American Society for Information Science and Technology, 56*(8), 883-888.

Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics, 1*(1), http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html.

Seaton, J. (2005). *Carnage and the media: The making and breaking of news about violence*. London: Allen Lane.

Smith, D. A. (2002). Detecting and browsing events in unstructured text. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 73-80).

Smith, S. (2005). Tapping the feed: In search of an RSS money trail. *Econtent, 28*(3), 30-34.

Sousa, V. D., Zauszniewski, J. A., & Musil, C. M. (2004). How to determine whether a convenience sample represents the population. *Applied Nursing Research, 2*, 130-133.

Swan, R., & Allan, J. (2000). Automatic generation of overview timelines. In E. Yannakoudakis, N. J. Belkin, M.-K. Leong & P. Ingwersen (Eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 49-56). Athens, Greece.

Tait, J. (2001). More Faust than Frankenstein: The European debate about risk regulation for genetically modified crops. *Journal of Risk Research, 4*(2), 175-189.

Thelwall, M., Prabowo, R., & Fairclough, R. (2006, to appear). Are raw RSS feeds suitable for broad issue scanning? A science concern case study. *Journal of the American Society for Information Science and Technology*. http://www.scit.wlv.ac.uk/%7Ecm1993/papers/Are%20rss%20feeds%20suitable%20for%20broad%20issue%20scanning%20preprint.doc

Thelwall, M., Vann, K., & Fairclough, R. (2006, to appear). Web issue analysis: An Integrated Water Resource Management case study. *Journal of the American Society for Information Science & Technology*.

Trammell, K. D., & Britton, J. D. (2005). *Gatewatching: The impact of blog content on the mainstream media.* Paper presented at the Internet Research 6.0: Internet Generations, Chicago.

Tsai, D. F. C. (2005). Human embryonic stem cell research debates: a Confucian argument. *Journal of Medical Ethics, 31*(11), 635-640.

Van Aelst, P., & Walgrave, S. (2002). New media, new movements? The role of the Internet in shaping the 'Anti-Globalization' movement. *Information, Communication & Society, 54*(4), 465-493.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. *Information Processing & Management, 40*(4), 693-707.

Viégas, F. B. (2005). Bloggers' expectations of privacy and accountability: An initial survey. *Journal of Computer-Mediated Communication, 10*(3), article 12. http://jcmc.indiana.edu/vol10/issue13/viégas.html.

Wayne, C. (1998). Topic Detection and Tracking (TDT): Overview & perspective. Retrieved February 14, 2006, from http://www.itl.nist.gov/iaui/894.01/publications/darpa98/html/tdt10/tdt10.htm

Wegrzyn-Wolska, K., & Szczepaniak, P. S. (2005). Classification of RSS-formatted documents using full text similarity measures. *Lecture Notes in Computer Science, 3579*, 400-405.

Wei, C. P., & Lee, Y. H. (2004). Event detection from online news documents for supporting environmental scanning. *Decision Support Systems, 36*(4), 385-401.

Wilt, J. (Ed.). (2003). *Making Humans: Mary Shelley, 'Frankenstein', H. G. Wells, 'The Island of Doctor Moreau'*. New York: Houghton Mifflin.

Wolpert, L. (2005). The Medawar Lecture 1998 - Is science dangerous? *Philosophical Transactions of The Royal Society B-Biological Sciences, 360*(1458), 1253-1258.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th international conference on machine learning (ICML 1997)* (pp. 412-420). Nashville, TN.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.