

project report for „Social Research and the Internet“

- Bloggers with an Agenda - Developing a Methodology to Assess whether Bloggers Rate Topics Independent from Media

Tobias Escher (tobias.escher@oii.ox.ac.uk)

Oxford Internet Institute, Hilary 2007

Abstract

This paper uses the theory of agenda setting to examine the relationship between blogs on the one hand and the media on the other. As previous research suggests that political topics on the blogosphere are mostly originating from the mainstream media the aim of this paper is to focus on the salience of issues on the respective agenda and to compare the importance of stories on the mainstream media with the importance of the same stories on the blogosphere. Specifically this research is interested in whether bloggers use relevance criteria that are different from the media or whether they adopt the issue salience assigned by media. A new methodology is proposed and piloted that applies different methods of web metrics to obtain empirical data that can help to answer the research question. Data is collected from Google News and posts in the blogosphere about the same topic are identified via Yahoo Term Extraction and Google Blogsearch. The rank of a story on the media agenda (according to number of articles for a story) is then compared with the rank of this story on the blogosphere agenda (according to number of posts for the same story). The results from the pilot study indicate that the proposed methodology can indeed be useful to analyze differences between media and blogging agenda. There are a number of methodological problems that are addressed. The general trend emerging from the data is that while both agendas are to some degree in tune, bloggers do not simply mirror the media agenda but apply different criteria to judge issue salience.

The continuously updated collection of data is made available via a web interface at
<http://uggeshall.adastral.ucl.ac.uk/blogagenda/index.html>

Table of Contents

Introduction.....	2
Previous Research.....	3
Hypothesis.....	5
Method	6
Findings from Pilot	12
Conclusion	16
Literature.....	17
Appendix.....	18

Introduction

There is much debate on the influence that the blogosphere or user-generated content in general has on traditional journalism and the mainstream media. Evidence suggests a growing importance of user-generated content with various established media outlets trying to integrate blogs and forms of citizen journalism into their portfolio (see BBC “email your picture” or Guardian’s “Comment is Free”). This debate must be seen in a wider context of traditional media critique that is questioning the role and power of mass media to define which topics are relevant (and more importantly, which ones are not) and to influence public opinion. Some have argued that blogs will be able to offer a counter to traditional media power by giving formerly passive audiences the technology to become producers of news themselves and to raise their topics to the awareness of a potentially huge audience on the Internet (Gorgura 2004). Some of these hopes seemed to be justified by a number of occasions where indeed stories that have not (or not sufficiently) been picked up by traditional media emerged on blogs and eventually gained so much momentum in the blogosphere that subsequently they had to be taken up by mainstream media. Examples include the case of US Senator Trent Lott who had to resign after remarks he made at a party (2004; Gill 2004), the debate about president Bush’s military credentials (Adamic and Glance 2005) or Salman Pax, the blogger from Baghdad.

In this way blogs did indeed overturn the traditional way of mass communication and empower the audiences. At the same time, the vast amount of content on blogs is either completely personal (Herring, Scheidt et al. 2004) or heavily citing traditional media

(Schmidt, Paetzolt et al. 2006). So the question persists how alternative the blogosphere really is and subsequently what can be its impact?

In this paper I treat blogs as an entity of its own, separate from the established mass media but also distinct from the general public as such – the audience of the news. Strictly speaking blogs are much more of a hybrid: They remove the distinction of active news producers and passive audience (Delwiche 2005) as part of the mass media audience forms the blogosphere. At the same time there is also an increasing incorporation of blogging into established media. Still as Halavais (Halavais 2002) points out, it has become fashionable to think of the millions of disparate blogs as an entity, the blogosphere and it has been shown that measured in structural terms, blogs do indeed increasingly resemble an interconnected space of its own (Kumar, Novak et al. 2003). I treat the blogosphere as a distinctive space that interacts with the media on the one hand and with the public on the other - a relationship that I outlined below.

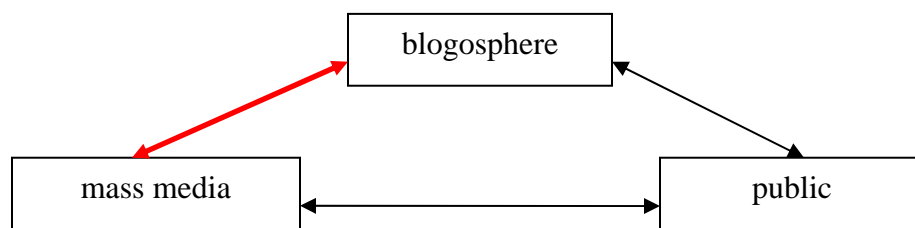


Figure 1: Simplified Model of Interaction between Media, Blogs and Public

In this paper, I will not look onto the public opinion side of effects but focus exclusively on the relationship of the blogosphere with the traditional mass media. This small scale research project is aimed at providing empirical data to help assess the relationship between blogs and established mass media. It will analyse the topics that are discussed in the mainstream media on the one hand and the blogosphere on the other hand by applying the theory of agenda setting and using methods of web metrics. By comparing the agenda of established media with the agenda of blogs I will try to analyse whether bloggers do attribute certain issues the same importance as the media does or whether they apply different criteria to judge their importance.

Previous Research

The theory of agenda setting originates from an older research that in the age of emerging mass media was enquiring about the effects of those media on the public, starting with Lippmann (Lippmann 1922). While the progressing research quickly moved away from the assumption of major media effects, the agenda setting approach of McCombs and Shaw (McCombs and Shaw 1972) came back to attributing the mass media rather

significant effects. However, this time effects were not defined as a change in affection (how people feel and think about a subject), but in cognition (what people know about a subject). Or, in the words of Cohen (Cohen 1963: 13), mass media "*may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to think about.*"

The main idea of McCombs and Shaw was that the media by reporting about a subject would put this on the agenda, this is in people's minds. By various means (e.g. according to number of times of coverage, placement, length etc.) they will also attribute these issues a certain importance. So as agenda they basically understood a set of issues with a certain rank order of importance or in the words of McCombs and Shaw: with a certain salience. The traditional agenda setting research usually follows a pattern outlined in Figure 2.

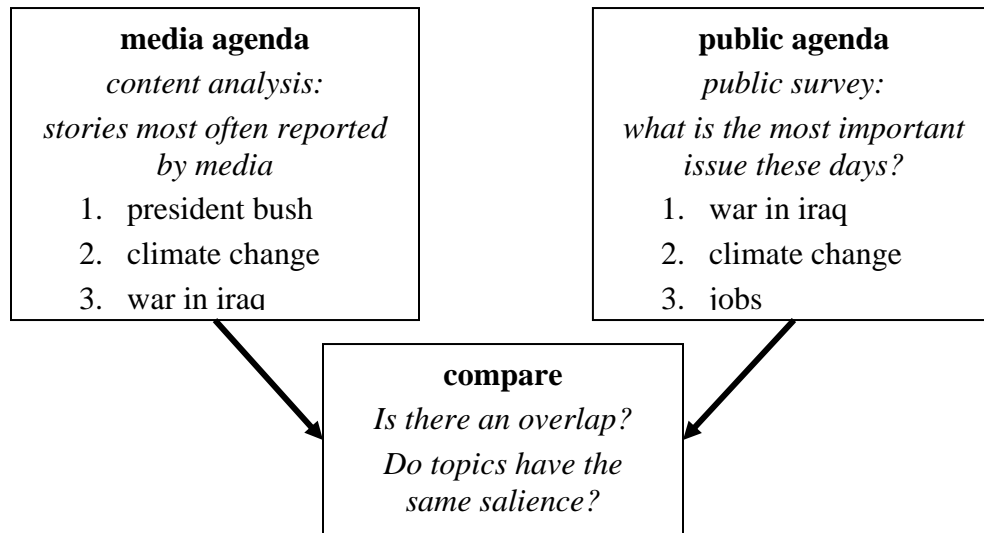


Figure 2: Methodology of Traditional Agenda Setting Research

McCombs and Shaw found a huge correlation of media and public agenda (almost 1.0 which in itself raises the question whether the method is actually valid). The question that all agenda setting researchers face then is one of causality: who is influencing whom? The question is of importance because as Dearing and Rogers (1996) point out that due to a limited attention span of the individual, agenda setting is usually a zero-sum game: new issues displace old ones. Who determines what issues feature high on the agenda has the power to shape public discourse (if not opinion). While the research that followed McCombs and Shaw found a variety of intervening variables (Erbring, Goldenberg et al. 1980; Rogers 1993), the overall conclusion of agenda setting research has been that there are effects from the media onto the agenda of their audience.

How does online communication and the tools that enable any individual to reach a global public with marginal cost change the rules of the game? Early studies of agenda setting and online communication have found clear evidence of an ongoing impact of traditional media (Roberts, Wanta et al. 2002). In recent years there have been a number of studies that explore how blogs fit into the theory of agenda setting, sparked by the previously documented examples of blog topics that crossed over in the mainstream media. Despite these examples so far researchers have usually come to the conclusion that instances of blogs setting the agenda of media are the exception rather than the rule. As Halavais points out, the major stories in blogs are usually related to the reporting in the media. Similarly Murley and Roberts (2006) in their content analysis of the top 20 US blogs (according to traffic) found heavy citing of traditional media (about 50%, the majority of those posts would not even comment on the cited source). Murley and Roberts found very little original reporting (6% of all posts) and therefore locate the role of blogs as interpreting and commenting on mainstream media rather than coming up with new topics. While Drezner and Farrell (2004) have noted that due to the speed of publication blogs can have a first mover advantage, they see the primary impact of blogs through media professionals reading them.

Hypothesis

As previous research overwhelmingly suggests that the media largely determines what topics are discussed on blogs this small scale project cannot search for evidence of the opposite. However, agenda setting is not only concerned with the kind of topics discussed but also with their salience. In other words, how important is an issue compared to others. The purpose of this study will be to establish the salience of issues on the media agenda as compared to the salience of those same issues on the blogosphere agenda. The question will therefore be:

Do bloggers assign salience to issues differently than traditional media outlets?

In this area there has been done little research. Thelwall and Hellsten's comparison of online discussions with media reports on the London bombings (Thelwall and Hellsten 2006) found some key differences between the two spheres. Similarly Delwiche (2003) carried out a study of about 800 blog posts that quoted news sources and points out that bloggers seemed to be relatively independent – one might even say uninterested – from the issues discussed in the media.

To put this question into the wider context of the initially outlined interaction between the media, blogs and the public: Should bloggers rate the importance of issues differently

from the media, then the blogosphere could indeed offer an alternative way of judging the news of the world for the wider public. In the words of Gorgura, blogs could “forge an online sphere of dissensus” (Gorgura 2004). However, it is still debatable whether blogs even when offering alternative or dissenting viewpoints can have an impact onto wider public debate or whether they rather facilitate Sunstein’s echo chamber (Sunstein 2001) by actually providing selected audiences with what they want to hear, not a broad spectrum of different opinions (Thompson 2003).

Method

Despite the number of studies that analyze blogs in relation to agenda setting none of them is convincing in methodological terms. Halavais (2002) simply uses word counts in blogs and does not conduct a proper comparison of blogs to the news agenda because the former seems to be just a mirror of the media. Delwiche (2005) has problems in operationalizing both the media agenda (simply using an Associated Press editor poll that asked “What was the most important story?”) and the public agenda (Gallup poll relying on a closed set of questions). Murley et al. (2006) would not compare the development of the stories over time but only look at a snapshot. Drezner et al. (2005) do an interesting study as they use a qualitative approach to measure the impact of blogs on the media (interviews with media professionals) but their conclusions rely on interpretation and assumptions and not on quantitative data.

This paper aims to contribute a new methodology to collect empirical data for the question under consideration. The methods that will be used mostly fall in the domain of web metrics that is usually concerned with counting entities or relations between entities on the web. While the application of web metric methods is prevalent in Computer Science (albeit under different names), there are examples of their useful application for research questions in the Social Sciences (Hindman, Tsioutsoulis et al. 2003; Thelwall and Hellsten 2006).

Given the topic of research, the key methodological questions to answer are:

- What is going to be the sample of media outlets and blogs?
- How can we establish the agenda (this includes salience of issues) of the media and of the blogosphere?
- How do we compare the two agendas?
- What is the timeframe (e.g. snapshot vs. longitudinal study)?

All of these questions are interconnected and have to be answered in relation to each other. For example, the methods and tools of data collection one chooses can largely determine the nature of the sample.

One of the key imperatives was to develop an automated approach. There are a variety of reasons for this that include easier data collection, the potential to use a much bigger sample and increased reliability as selection is based on certain criteria instead of subjective decisions by the researcher/coder. However, despite the named benefits there is a rather large initial effort required in order to set up the technology. What is more, the reliance on mainly technological means for collecting and analyzing the data has pitfalls which will be outlined later.

Sample Selection

The basic decision to take is which languages (and therefore countries) to include and whether or not to focus on specific topics. For methodological reasons (see below) I decided to focus on the global agenda of English speaking news on the world in topics we might term “serious”. The topics are selected from the Google News’ “World” category¹ which excludes sports, entertainment and news that are only relevant to the US. Such a wide sample brings up immediately the issue of how homogenous the media (or blogosphere) agenda is worldwide. It would be desirable to conduct a comparison on state level and even on this level, “the explosion of media inputs erodes the notion of a unified media agenda” as Chaffee & Metzger, 2001 point out. However, it is difficult to establish to which country a blog belongs (or what is intended audience is) and therefore to establish an agenda of a specific “national” blogosphere. While one solution could be to select a number of clearly assignable blogs for analysis this would immediately create the next difficult problem of how to select representative blogs. Choosing such an inclusive sample is therefore necessarily homogenizing much of the diversity of the media but will hopefully provide a bird’s eye view of the global media and blog landscape.

Establishing a Media Agenda

One of the advantages of the Internet is that enables access to the whole variety of media outlets as virtually all of the important ones are available online. The global Google News service available from <http://news.google.com/> is a news aggregator that promises

¹ see <http://news.google.com/?ned=us&topic=w>

a convenient access to the English speaking news of the world. It automatically checks the content of 4,500 English language news sources (not all of them are traditional in our sense) from around the world² and automatically groups different news articles together if the computer-algorithms determine that they seem to belong to the same story. For the purpose of this project, I will consider the number of articles Google reports for a certain story as an indicator of the salience of this story such that many articles signal a high importance of that story on the media agenda. Google News also provides standardized XML Atom feeds that can be used to comfortably collect the data by computer scripts. Please refer to the technical Appendix for an in-depth outline of the technical setup. While there are other news providers online, none of them offers the clustering of news articles together with the easy option for interfacing with the data.

As previously explained the Google News “World” category is used which is updated several times per hour³ and lists the most important stories in this category. This list is continuously collected every twenty minutes together with the first ten articles Google offers for each story. The data collection should give a pretty complete picture of each story from a variety of viewpoints as well as provide the foundation for a longitudinal study. For each day, stories will be ranked by the number of articles. In divergence from traditional agenda setting research, the unit of analysis is therefore stories, not issues. Several different stories will relate to the same issue, such as the war in Iraq. It would be possible to use the proposed methodology and manually group stories under broader issue categories. This is beyond the scope of this paper but I argue that there is still enough leverage in comparing the ranking of different stories as it will still enable to assess difference between journalists and bloggers in attributing salience to certain stories that relate to a range of issues.

There are a number of problems relating to the use of Google News: As often in search engine results, the number of articles for a story that Google reports tends to be inflated. Another problem is the fact that the clustering method of Google is not publicly described. We simply do not know exactly how Google groups different articles to one story and have no control over it. For the same reason it is not explicitly clear what articles fit in the World category. This has implications for the reliability of the research as Google may alter algorithms that subsequently lead to different results.

² http://news.google.com/intl/en_us/about_google_news.html [Accessed 02.04.07]

³ <http://www.google.com/support/news/bin/answer.py?answer=2828&topic=8868> [Accessed 02.04.07]

Establishing a Blogosphere Agenda

Despite the existence of a variety of blog aggregators there is unfortunately no comparable service to Google News available for the blogosphere. There is no transparent way of identifying top stories on the blogosphere. One option could be to collect data from blogs directly (see for example the Mozdeh project of Mike Thelwall at <http://mozdeh.wlv.ac.uk/>) but it is a huge effort, it takes time to build a corpus and it again brings up the problem of sample selection. To compare ranking differences in assigning salience to stories, I could start by analyzing how strongly the stories discussed on the media do feature on blogs. While this does put the emphasis exclusively on the media for bringing up stories, this seems permissible given the findings from previous research as well as considering the scale of this project.

To find out how good the stories from the media agenda do on the blogosphere one could identify the number of posts that talk about a story from the media agenda. I argue that the more posts discuss a story the higher is the stories' salience on the blogosphere. To identify media stories on blogs key terms are extracted from the articles relating to a story and then used to search the blogosphere for the number of times blog posts mention these key terms. While statistics are difficult, search of the blogosphere is rather comfortable.

The identification of relevant keywords is a big computer science problem and also related to methods of quantitative content analysis. For this research I use the Yahoo Term Extractor⁴ that offers a simple and easy way of extracting significant words and phrases from a text via automated means. The immediate problem, similar to Google News clustering, is that we do not know how exactly it is being done. Nevertheless, initial tests (see below) show that it works reasonably well.

For this research the terms of the first ten articles are extracted for each story on Google News and ranked according to number of times they appear in these articles. The underlying assumption is that the phrases that are most often quoted in the articles should give a pretty good description of the story. The two most common phrases are then subsequently used to search for the number of blog posts that mention both of these terms. Here Google's blog search (<http://blogsearch.google.com/>) is employed which also allows automated collection of search results via news feeds.

One problem with this approach is that the use of language in media and on blogs might be different. Blog posts concerning the same story might use a vocabulary very different

⁴ <http://developer.yahoo.com/search/content/V1/termExtraction.html> [Accessed 02.04.07]

from the media and therefore will not show up by searching with keywords obtained from the media texts (for example a news agency might report about “insurgents” while a blog might refer to “freedom fighters”). This all indicates that further research would be needed to establish the ideal selection of keywords for searching blogs.

Roberts et. al (2002) crucially point out the time lag between reporting in the news and discussion online. We must assume that it might take some time before bloggers take up news events that have occurred. Fortunately the search function of Google blog search allows specifying time ranges. For this research the search is limited to posts that contain the keywords and occur not more than two days before and one day after the story featured on Google News. The timely connection to the news story should also help to return more relevant posts. Again there is scope for further research on what the best time frame is.

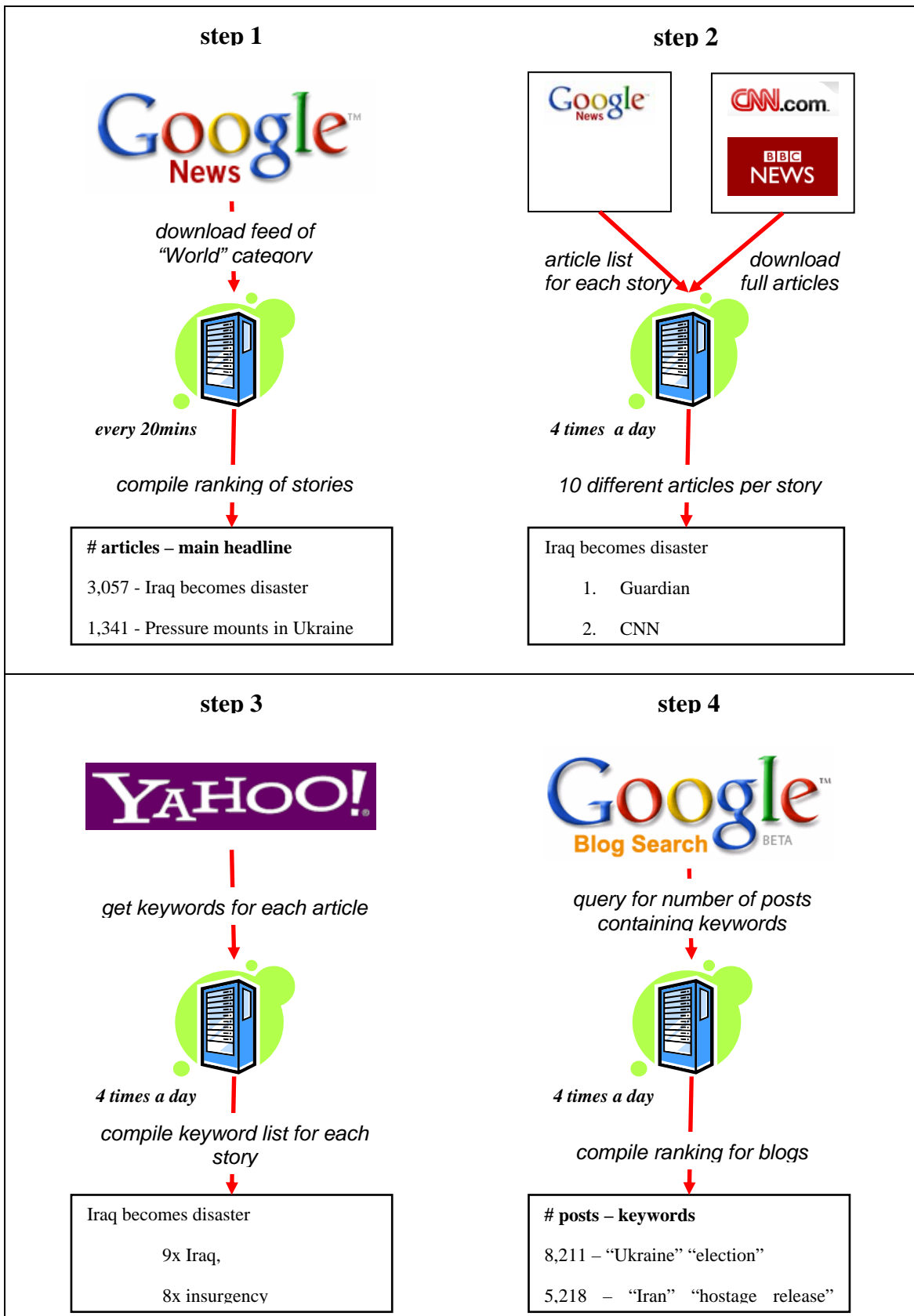
There are a number of issues concerning the use of Google blog search: The returned results tend to include a rather high number of sources we might call traditional (such as media sites), in that sense Google’s criteria of what constitutes a blog is not quite clear as well as coverage of the index and the update interval. A further problem is how the keywords are treated internally. One would assume that keywords are stemmed but tests with similar keywords (e.g. “weapon” or “weapons”) deliver different results.

Comparing Agendas

The previous sections have outlined the means of data collection together with their pros and cons. Figure 3 gives a graphic overview of the data collection process. However, once the data is collected, how should the agendas be compared?

The approach used in this paper is for each day to rank order the stories on each agenda according to the number of articles and number of posts respectively using an ordinal ranking. This allows to state relative importance of the stories compared to the other stories on the agenda. However, at the same time it does simplify the question of salience as it uses only an ordinal measure (the rank) instead of the metric measure (post/article count). In future one could also make use of the total numbers in order to assess differences in salience in more detail.

Figure 3: Methodology: Overview of Data Collection



The analysis should clearly focus on differences in ranking. In order to ease this a simple measure is calculated that reports the normalized differences between the two rankings for each story by subtracting the blog rank from the media rank and dividing by the total number of stories (ie. ranks). This yields a comparative measure of relative importance of the story on the blogging agenda as compared to the media agenda: negative values indicate a story is less important on blogs, positive values indicate a story is more important on blogs. In addition, the root mean square (RMS, see Appendix) of all individual differences is computed to deliver a single value that assess the overall overlap of the two agendas with values closer to 0 indicating high overall overlap, values closer to 1 high degrees of difference.

The data for the analysis can be comfortably accessed via a web interface (available at <http://uggeshall.adastral.ucl.ac.uk/blogagenda/index.html>) that is illustrated in Figure 4. This interface can then be used by the researcher to establish whether certain stories tend to be more often ranked lower or higher on the blogosphere.

Ethical Considerations

There are no ethical real ethical issues involved as all the data is publicly available on the Internet and the analysis focuses on aggregates of news articles and blog posts. What is more, data is only accessed from big information providers that explicitly make the information available to the public. In addition the amount of data is comparatively small therefore bandwidth consideration do not apply (Thelwall and Stuart 2006).

Findings from Pilot Study

Data has been continuously collected since 4th April 2007. This serves as a pilot of the methodology that allows to for its assessment as well as provides two snapshots of the data for analysis in relation to the research question.

Methodological Issues

The first point of interest is to assess how good the keyword selection is. On the positive side it turns out that keywords do not always have to be very specific. Due to the time frame for the blog search even vague keywords deliver relevant results. For example "interview" and "confiscated" will correctly bring up many blog posts related to the headline "*Jordan confiscates videotape of interview with Prince Hassan*".

Query Media and Blogosphere Agenda: 46 stories for 2007-04-26

Select a date and click submit to view data for this day!

(Blog agenda is constructed with time lag of about two days. Access more recent information on news stories and their keywords via navigation bar)

2007-04-26 order by: mediarank blogrank difference

root mean square of normalized difference: 0.320

Mediarank	Blogrank	normalized difference (as bloginterest)	Title (link to stored articles for this story)	Keywords (link to all keywords)	Number of Articles on Google News	Number of Posts on Google Blogsearch (link to Blogsearch)
9	40	-0.674	Mogadishu war enters second week, no let-up in sight - Reuters AlertNet	"mogadishu somalia" "ethiopian troops"	622	60
14	34	-0.435	Afghan soldiers die in bomb attack - Aljazeera.net	"afghan" "provincial police"	409	102
21	41	-0.435	Russia waiting for UN resolution on Kosovo - International Herald Tribune	"serbia" "ethnic albanian majority"	288	57
10	29	-0.413	Mexico defies Church on abortion - Times Online	"mexico city" "12 weeks of pregnancy"	610	133
26	44	-0.391	Syria's ruling party wins in vote - Houston Chronicle	"baath party" "parliamentary elections"	141	16
8	23	-0.326	Nigerian opposition calls for peaceful protests - Canada.com	"umaru yar adua" "opposition"	691	226
30	45	-0.326	Mazud to decide whether to probe Olmert - Ynetnews	"state comptroller" "trade minister"	126	13
22	36	-0.304	Journalists rally for release of abducted BBC reporter - Xinhua	"alan johnston" "palestinian journalists"	251	85
1	14	-0.283	World leaders bid farewell to Yeltsin - Indian Express	"boris yeltsin" "laid to rest"	3431	408

Figure 4: Screenshot of Web Interface

However, clearly not all keywords were good descriptions. As one of the reasons emerged the fact that from the ten articles downloaded from Google News, many were actually very similar or even the same as different media would report the same news agency article. To improve quality the program was modified to collect only articles that have a distinct title and only one article per medium. In addition the number of downloaded articles was increased to 15 in order to increase variation which seemed to immediately improve the keywords. The initial results suggest that the two query terms for the blog search are usually sufficient to deliver relevant results. As the blog search will often deliver only little results, further addition of keywords would probably lead to more empty searches.

However, there is still the problem of how to best select the keywords to use. For example, for the French elections we get 13x "nicolas sarkozy", 12x "segolene royal" and "12x francois bayrou". It is hard to decide by automated means which to use. At the

same time, some of the extracted terms are general words such as “Associated Press”. In order to improve the selection in the future one could create a black list that removes some of these terms. Other means of improving keyword selection could make use of the mentioned “term frequency–inverse document frequency”⁵ statistics, stemming of keywords as well making sure that the two query terms are sufficiently different (for example not “*afghanistan taliban*” and “*taliban afghanistan*”). Keyword selection could also make more use of the rank of the keywords, for example it might make sense to only take the top 5 keywords for each article and count their occurrences. There is also the option of selecting keywords manually and use the keyword extraction as a tool to assist this process. However, the necessary effort (e.g. multiple coders to ensure inter-coder reliability) is rather big.

Another problem that has arisen is that some stories reported by Google News are actually rather similar. These are covering the same topic but with different foci. It is basically impossible to capture these minor differences by only two describing key phrases. Consequently this results in some of the blog searches basically using the same keywords (in this case “iran” and “british soldiers”).

Summarizing the keyword selection problem, the main question is just how much error can be justified by the amount of data that can be processed by automatic means in contrast to a better precision but less data produced manually.

Results

For this research the 46 and 54 stories collected for the 26th and 27th April 2007 respectively are analyzed.

26th April 2007

The average RMS is 0.320 which indicates some difference between the agendas. There are about a third of all stories that feature about equally important on both agendas, identified as a normalized difference of no more than 0.1. However, the only one of these that is a major story (top 10) is the French election. The other stories with similar salience on both agendas tend to be less important. A closer examination of the stories with huge ranking differences reveals that these differences are often due to keyword selection. The keywords used for searching blog posts are either too specific (yielding little results and therefore low ranking) or too general (yielding many results and therefore high rank). So

⁵ see Wikipedia for an introduction to TF-IDF: <http://en.wikipedia.org/wiki/Tf-idf> [Accessed 27.04.07]

with this in mind each story has to be assessed individually whether the queried keywords do actually reflect the story. If so, then the ranking differences are meaningful.

A very good example is the 70th anniversary of the Nazi bombings of Guernica. The query looks for posts containing "*pablo picasso*" and "*guernica*" and is therefore correct. While this story features very low on the media, it is much more important on the blogosphere. Even better, we see that the visit of the Japanese Prime Minister Shinzo Abe generates huge discussions in the blogosphere but relatively little on the media. The same applies to Chinese concerns over nationals killed in Ethiopia and the discussion about the Dalai Lama. Conversely bloggers seem largely uninterested in Syria and North Korea. Still it seems like regional topics are more salient on blogs, maybe because in non-English speaking countries there is a greater share of English language blogs than English language media.

27th April 2007

A reasonable overlap as indicated by the RMS of 0.31. Again, many stories (~40%) have little ranking difference and are less important but these include also a top5 story (Russia and missiles). As one might have expected, Amnesty International's new report on death penalty generated much more publicity on the blogosphere than on the media as did another regional topic, the agreement between Myanmar and North Korea. The latter is interesting as the previous day the agreement between North Korea and Syria did not feature prominently on blogs, maybe indicating a more Asian centric blogosphere. Following up from the day before, the anniversary of the bombing of Guernica is still more important on blogs.

In contrast, the death of Russian cellist Mstislav Rostropovich, the UN ambassador's visit to Kosovo and Mexico City's plans to allow abortion do not generate much discussion online. A little bit surprising might be that Chinese plans to enhance environment conservation do not find a bigger echo in the community of bloggers.

An interesting case is the story about a military helicopter crash in Chechnya which appears twice on the agenda: once it is more important on blogs and once more important on the media. The blog search is executed twice, each time with keywords extracted from a different set of articles and consequently yielding different results. This highlights again the need for improvement in keyword selection as well as some qualified interpretation of the data.

Conclusion

In this paper I have proposed and piloted a web metrical method of researching about the relation of the media and the blogging agenda, in particular about ranking differences between the two spheres. As the method employs a variety of proprietary information services and different tools of data collection there are numerous potential pitfalls that have been addressed. Some relate to the quality of data available via the Web and the opaque methods by which it is provided from Google and Yahoo, others relate to the researcher's decisions about how to use and code the data (for example the number of query terms for the blog search, the time span for which to search blogs etc.). Nevertheless the pilot demonstrated that the proposed method can indeed offer useful data to compare the agendas of media and blogosphere. Due to the scope of this research the pilot study could only analyze two snapshots from the collected data but the results indicate that while there is a considerable overlap in ranking stories by no means do blogs just mirror the media agenda. Therefore bloggers do indeed seem to apply different criteria to assign salience to issues and related stories.

Future research should include a longitudinal study of how the salience of stories develops over time and also how stories themselves change. The tools that have been prototyped for this research offer a good opportunity to pursue this goal. In addition future research should try to establish ways on collecting topic rankings from blogs independent from media stories in order to see whether the blogosphere can establish alternative topics and who leads or lags the agenda.

Literature

- Adamic, L. and N. Glance (2005). The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.
- Cohen, B. C. (1963). The press and foreign policy. Princeton, N.J., Princeton University Press.
- Dearing, J. W. and E. M. Rogers (1996). Agenda-setting. Thousand Oaks, Calif. ; London, Sage.
- Delwiche, A. (2005). "Agenda-setting, opinion leadership, and the world of Web logs." First Monday **10**(12).
- Drezner, D. W. and H. Farrell (2004). The Power and Politics of Blogs. American Political Science Association.
- Erbring, L., E. N. Goldenberg, et al. (1980). "Front-Page News and Real-World Cues: A New Look at Agenda-Setting by the Media." American Journal of Political Science **24**(1): 16-49.
- Gill, K. E. (2004). How can we measure the influence of the blogosphere? World Wide Web Conference, New York.
- Gorgura, H. (2004). The War on the Terror Consensus: Anti-War Blogs as an Online Sphere of Dissensus. Internet Research 5.0, Sussex, Association of Internet Researchers.
- Halavais, A. (2002). Blogs and the "Social Weather". Internet Research 3.0: Net / Work / Theory, Maastricht.
- Herring, S. C., L. A. Scheidt, et al. (2004). Bridging the Gap: A Genre Analysis of Weblogs. 37th Hawaii International Conference on System Sciences, Hawaii.
- Hindman, M., K. Tsioutsouloukalis, et al. (2003). Googlearchy: How a Few Heavily-Linked Sites Dominate Politics on the Web. Annual Meeting of the Midwest Political Science Association, Chicago, IL.
- Kavanaugh, A., J. M. Carroll, et al. (2005). "Community Networks: Where Offline Communities Meet Online." Journal of Computer-Mediated Communication **10**(4): 00-00.
- Lippmann, W. (1922). Public opinion. New York, Macmillan.
- McCombs, M. E. and D. L. Shaw (1972). "The Agenda-Setting Function of Mass Media." The Public Opinion Quarterly **36**(2): 176-187.
- Murley, B. and C. Roberts (2006). Biting the Hand that Feeds: Blogs and second-level agenda setting. Southern Political Science Association, Atlanta.
- Roberts, M., W. Wanta, et al. (2002). "Agenda Setting and Issue Salience Online." Communication Research **29**(4): 452-465.
- Rogers, E. M. (1993). "The anatomy of agenda-setting research." Journal of Communication **43**(2): 68.
- Schmidt, J., M. Paetzolt, et al. (2006). Stabilität und Dynamik von Weblog-Praktiken? Ergebnisse der Nachbefragung zur "Wie ich blogge?!"-Umfrage. Berichte der Forschungsstelle „Neue Kommunikationsmedien. Bamberg, Germany.
- Sunstein, C. R. (2001). Republic.com. Princeton, N.J. ; Oxford, Princeton University Press.
- Thelwall, M. and I. Hellsten (2006). "The BBC, Daily Telegraph and Wikinews timelines of the terrorist attacks of 7th July 2006 in London: a comparison with contemporary discussions." Information Research **12**(1).
- Thelwall, M. and D. Stuart (2006). "Web crawling ethics revisited: Cost, privacy, and denial of service." Journal of the American Society for Information Science and Technology **57**(13): 1771-1779.
- Thompson, G. (2003). "Weblogs, warblogs, the public sphere, and bubbles." Transformations(7).

Appendix

Technical Notes

The system uses exclusively Free Software components. It consists of a MySQL database that is filled with data via Perl scripts. The content is then served to a publicly available website, using the Apache web server. Database administration has relied on phpMyAdmin.

Google News

Every 20 mins a Perl script downloads the current Atom feed of the Google News “World” category. The information is parsed and inserted into a MySQL database. Should during one day the same story come up several times, it is only stored once but with the highest number of results reported from Google. Twice a day another Perl scripts accesses this information to download the first 15 articles for each story. Considered are only articles that have a unique headline and only one article per media outlet. A special Perl module is used (HTML::ContentExtractor) in order to extract only the relevant text of the downloaded web pages that is the content of the articles and not the advertisements etc. On demand a Perl script sends the text of each article to the Yahoo Term Extraction tool (with the optional query parameter set to the title of the article) which will respond with a list of most important terms/phrases that are subsequently stored in the MySQL database. A SQL query will then deliver a list of how often a term is mentioned by the articles for each story (so this could be a maximum of 15 times and a minimum of one time if only one article uses this word deemed important by Yahoo term extraction).

Blog Search

From the ranking of keywords for each story, the two phrases (could be one or more words) most often mentioned will be used by a Perl script that queries the Google Blog search⁶ several times a day for all blog posts that contain the two phrases. Additionally, the query is limited to a time span of four days, starting two days before and stopping one day after the story appeared on Google News. The resulting count of posts is stored in the MySQL database.

⁶ <http://blogsearch.google.com>

Comparing Media and Blogosphere Agenda

A website (accessible from <http://uggeshall.adastral.ucl.ac.uk/blogagenda/index.html>) can be used to compare the agendas of the two spheres as well as to access detailed information on the collected articles and extracted keywords. For every day, the stories from Google News will be shown, once ranked according to number of articles on Google News, once ranked for number of posts containing the corresponding keywords. The difference between the rankings is calculated and normalized such by dividing the difference through the total number of ranks. This yields a difference normalized to be in the range of -1 to 1. As an overall measure of difference between the two agendas the root mean square is calculated as follows:

$$x_{\text{rms}} = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \dots + x_n^2}{n}}$$

Figure 5: Calculation of Root Mean Square
(picture obtained from Wikipedia article on that subject)

This yields result to be in the range of 0 to 1. Smaller values indicate little difference (this is high concurrence) while conversely higher number signal major differences in between both rankings.