The TREC Blogs06 Collection : Creating and Analysing a Blog Test Collection

Craig Macdonald, Iadh Ounis
Department of Computing Science
University of Glasgow Scotland, UK
{craigm,ounis}@dcs.gla.ac.uk

ABSTRACT

The explosion of blogs on the Web in recent years has fostered research interest in the Information Retrieval (IR) and other communities into the properties of the so-called 'blogsphere'. However, without any standard test collection available, research has been restricted to unshared collections collected by individual research groups.

With the advent of the Blog Track running at TREC 2006, there was a need to create a test collection of blog data, that could be shared among participants and form the backbone of the experiments. Such a collection should be a realistic snapshot of the blogsphere, of enough blogs as to have recognisable properties of the blogsphere, and over a long enough time period that events should be recognisable. In addition, the collection should exhibit other properties of the blogsphere, such as splogs and comment spam. This paper describes the creation of the Blogs06 collection by the University of Glasgow, and reports statistics of the collected data. Moreover, we demonstrate how some characteristics of the collection vary across the spam and non-spam components of the collection.

Keywords

Test Collections, Blogs, Feeds, Splogs, Spam

1. INTRODUCTION

The rise of blogging on the Internet - the creation of journal-like web page logs - has created a highly dynamic and interwoven subset of the World Wide Web that evolves and responds to the real-world events [10]. The size of the so-called blogsphere (or blogosphere) - which is the collection of blogs on the Internet - has been growing exponentially for the last three years, with the number of blogs tracked by Technorati doubling every six months [15].

The growth of the blogsphere has led to the creation of new search engines tracking and searching the contents of blogs, thus servicing the need of Internet users for effective retrieval tools for the blogsphere. Today, there exist several blog search engines - some focusing on searching blogs, such as BlogDigger, BlogPulse and Technorati; and some specialised services from the main Web search engines, such as Google, Yahoo! and AskJeeves.

The need for blog search engines that are separate from mainstream Web search engines is motivated by the fact that the use of blog search engines differs from the use of conventional Web search engines. In [4], Broder designated three types of information needs in Web search: informational, transactional, and navigational. However, in [12], Mishne & de Rijke found that the information needs in blog search differ substantially. Transactional queries make less sense in blog search as people do not buy commercial products on the blogsphere. There is less evidence of navigational queries, as conventional Web search engines are suitable for this task. In fact, most queries are of an informational nature. A large number of queries are of a repetitive nature - caused by automatic searches by end-user tools - to identify new and timely articles about a general interest. This showed a prevalence of filtering queries not observed in Web searching studies. The remaining queries are adhoc - the user is looking for blogs with an interest in a topic (called Concept queries), or blog posts which discuss a (named entity) topic, called Context queries. Between these types of queries, the major aspects of the blogsphere are covered: the temporal and the discussive, opinionated nature of posts [14].

The TREC Blog track was proposed in October 2005, with the central aim of supporting research into retrieval in the blog search context, in particular into the major features of the blogsphere: the temporal aspects, and the opinionated nature of posts. The initial task of the TREC 2006 Blog track is the Opinion task. Participants are asked to retrieve blog postings about a (named entity) topic, but also express an opinion about the topic[1].

Generally, in a TREC track, participant research groups participate in tasks, using the same corpus and topics. With the creation of a track at TREC, comes the need for shared resources for experimentation. An IR test collection consists of an initial corpus of documents, a set of encapsulated information needs (topics), and a set of relevance judgements. However for the TREC 2006 Blog track, there was not a suitable corpus of blog documents available to distribute to the participants. Such a collection should be a representative sample of the blogsphere, suitable for the envisaged task of the track, but also suitable for other unenvisaged experiments using blogs.

The remainder of this paper describes the design and creation of the Blogs06 TREC test collection[2], and examines the statistics of the collected data. Section 2 discusses appropriate properties that a blog test collection should ex-

[1]For more information about TREC 2006 and the Blog track, see `http://trec.nist.gov/tracks.html`

[2]The Blogs06 TREC test collection can be obtained from `http://ir.dcs.gla.ac.uk/test_collections/`

Department of Computing Science, University of Glasgow Technical Report
 TR-2006-224.

hibit. In Section 3, we discuss the three phases of creating the Blogs06 test collection. Section 4 provides an overview of the statistics of the collection, and analyses in detail the use of dates and times in XML feeds, across the spam and non-spam components of the collection. Section 5 assesses the coverage of the PubSub ping log over the collection. Section 6 examines some term features of the spam and non-spam documents, while Section 7 investigates the linkage structure of the collection, with respect to the spam and non-spam documents. We summarise some related work in Section 8, and provide concluding remarks in Section 9.

2. DESIRED CORPUS PROPERTIES

In this section, we describe the properties desired of a suitable test collection for blog research. These properties are supported by the motivations described in Section 1. To create a realistic setting for blog search experimentation, the corpus should reflect characteristics of the real blogsphere.

A feature of blogs is that they have XML feeds describing recent postings. Two competing formats for feeds are prevalent in blogs: RSS and Atom. As a real blog search engine would need to cope with both types of XML feed, we chose not to restrict our collection to either format alone.

While RSS and Atom both provide an area for content, around 30% of feeds do not include the full textual content for each post [11]. Additionally, the vast majority of blogs allow comments to be added to postings by readers, but the comments are not included in the feed. The presence of comments allows researchers to see what readers think about a post, or also to see blog comment spam. For these reasons, we chose to save both XML feeds and HTML permalink documents in our collection. This would facilitate studies into how useful the HTML content is over the XML feed alone. Additionally, we save the homepage of each blog at the time each feed is collected.

In terms of breadth, we set out to collect blog postings over a period of time. The time span of the collection should be long enough to allow filtering, topic detection and event tracking experiments to take place.

Given the substantial time period desired, it was unfeasible to track every blog on the blogsphere, or even every English blog, while still keeping the collection at a size that could be easily redistributed. We chose to monitor about 100,000 blogs of varying quality, which should be a representative sample of the blogsphere at large. While mainly in English, the collection should contain some blogs in non-English languages, and a significant amount of spam blogs (splogs), to mimic the problems faced by blog search engines.

In the next section, we describe the three phases of creating the Blogs06 collection.

3. CORPUS CONSTRUCTION

The corpus construction for the Blogs06 collection lasted four months, which can be broken down into several stages: firstly, the selection of suitable blogs to crawl; secondly, fetching the appropriate content from the Web; and thirdly organising the collection into a reusable form.

The following sections describe each phase of the corpus construction in further detail.

3.1 Blog Selection

The Blogs06 test collection differs from the standard Web test collections in that no new blogs were added to the collection after the first day of the crawl. The blogs to be included in the collection were pre-determined before the outset of the fetching phase. In total, we selected 100,649 blogs for the Blogs06 collection. These came from several sources:

- **Top blogs (70,701):**

  To form a usable test collection, we aimed to include top blogs from the Web. A list of blogs, which included a sample of top blogs[3], was provided by a specialised blog search company, via the University of Amsterdam.

- **Splogs (17,969):**

  Splogs are a large problem on the blogsphere, and blog search engines are faced with the growing problem of identifying and removing spam blogs from their indices. Splogs are generated for two overlapping motives [8]: Firstly, fake blogs containing gibberish or plagiarised content from other blogs or news sources host profitable context based advertisements; Secondly, false blogs are created to realise a link farm intended to increase the search engine ranking of affiliated sites.

  A list of known spam blogs was also included in the test collection. The spam component forms a reasonable component of the collection, such that participants are faced with a realistic scenario.

- **Other blogs (11,979):**

  Finally, we supplemented the collection with some general interest blogs, such as news, sport, politics (US & UK), health etc. These additional blogs were found by manual browsing of the web or of sites and blogs relevant to the corpus purpose, and were added to give a variety of genres of material in the collection, and to ensure that there was content in the collection that would be readily accessible to TREC assessors.

3.2 Fetching Content

The content of the Blogs06 collection was fetched over an eleven week period from the 6th December 2005 until the 21st February 2006. Fetching the content from the blogs over the period was broken down into two tasks: regularly fetching the feeds and homepages of each blog; and fetching newly found permalinks that were extracted from the feeds. These were known as the Feeds and Permalinks crawls respectively. These are described separately in the following sections.

3.2.1 Feeds Crawl

Ideally, we wanted to identify as much new content from each of the blogs as possible over the entire period of the collection. We desired to check the feed of each blog once a day. However, because as much as 35% of the collection originated from Blogspot.com, we did not wish to poll the Blogspot servers for 35,656 feeds and 35,656 homepages each day. Doing so would have meant sending requests to the Blogspot.com servers at a rate of around one request per second - the required rate to complete all 71,312 requests in a 24 hour period. Although the Blogspot servers can no doubt

[3]We were not informed as to the way in which the top blogs were determined.

handle such load, to do so would have been considered a breach of the politeness protocol, and we would have run the risk of being banned from connecting to Blogspot servers.

Instead, we opted to poll each feed once a week. The set of feeds was broken down into 7 similarly sized bundles, one for each day of the week. Feeds from each of the large components, namely Blogspot, Livejournal, Xanga and MSN Spaces, were evenly distributed across the 7 bundles.

Each time a feed was downloaded, the homepage of the blog was extracted, as were the URLs of all the permalink documents. The homepage was added to the queue of URLs to be fetched that day, while the permalinks were written to a file on disk, for later fetching. The time delay between each request to a given IP address was 2 seconds. This meant that the feeds and homepages crawl typically finished in 5 hours each day. Furthermore the crawler abided by all of the robots exclusion protocols existing [9, 1]. This meant that homepages or permalinks documents linked to in feeds may not be available in the collection itself, as they have been explicitly disallowed by the exclusion protocols.

3.2.2 Permalinks Crawl

As discussed in Section 2, we desired the fetched permalink documents (i.e. HTML blog posts) to include comments left by readers and also any possible comment spam. If the permalink document was collected as soon as the permalink URL was discovered, the comments may not have been left. Instead, we delayed fetching newly found permalinks for at least 2 weeks. After an initial 2 week delay from the start of the feed crawling, we started collating the permalink URLs extracted by the daily feed crawler, removing duplicates, and fetching the permalink documents. As each feed could generate links to many new permalink documents, the permalinks crawl for a week's worth of new permalinks URLs could take more than one week to complete. For instance, there were 322,692 Blogspot permalinks found in the first week of the crawl. At one fetch every 2 seconds, these permalinks took 8 days to collect.

3.3 Organising the Collection

Once all crawled data was collected, we had to reorganise the collected data in a format easy to use for research purposes. We aimed here to adhere to the general layout formats from preceding TREC Web collections, as this would allow participating groups easier reuse of existing tools. In the Blogs06 collection, we collected the feed and homepages for each blog multiple times, and each newly found permalink document once. Because of this inherent structure between these different types of data, we supplemented the traditional TREC format with additional 'tags', to show the linkage between the different components of the collection. Figure 1 shows the format of one feed from the Blogs06 collection.

The collection was organised in a day-by-day format, one directory for each day of the collection. For each day, the feeds, homepages, and permalink documents were placed in separately named files. Each feed, homepage, and permalink document were given unique identifiers. In the case of feeds and homepages, these unique identifiers were the same throughout multiple fetches over the period of the collection. A DOCNO uniquely identifies one permalink document. From the DOCNO, it can be determined what day the permalink URL was first discovered, what file number the document is stored in, and the offset within the file. Note that the two week delay in fetching the permalink documents is hidden from research users of the collection, making it easier for them to associate documents with the dates on which they were found.

```
<DOC>
<FEEDNO>BLOG06-feed-001002</FEEDNO>
<FEEDURL>http://www.henrikbennetsen.com/wp-rss2.php#</FEEDURL>
<BLOGHPNO>BLOG06-bloghp-001002</BLOGHPNO>
<BLOGHPURL>http://www.henrikbennetsen.com/#</BLOGHPURL>
<PERMALINKS>
http://www.henrikbennetsen.com/?p=85#
 BLOG06-20051206-012-0001942855
http://www.henrikbennetsen.com/?p=83#
 BLOG06-20051206-012-0001954556
</PERMALINKS>
<DOCHDR>
http://www.henrikbennetsen.com/wp-rss2.php# 200512663735 25595
Date: Tue, 06 Dec 2005 20:37:26 GMT
Server: Apache
Content-Type: text/xml; charset=UTF-8
Last-Modified: Tue, 06 Dec 2005 18:55:22 GMT
X-Pingback: http://www.henrikbennetsen.com/xmlrpc.php

</DOCHDR>
<?xml version="1.0" encoding="UTF-8"?>
<!-- generator="wordpress/1.5" -->
<rss version="2.0"
...
</DOC>
```

Figure 1: Sample markup of an RSS feed from the Blogs06 collection. Standard TREC tags are present, such as DOC, DOCHDR, as well as supplemental tags devised for the Blogs06 collection.

Algorithm 1 : Building the collection

```
Assign a unique identifier for each feed
For each day:
   For each feed in day:
      For each permalink URL in feed:
         Normalise the URL
         Apply redirects to the URL
         Discard if URL already processed
         Assign document number to URL
         Find and write out permalink document
      Write out feed and homepage
```

Algorithm 1 details the algorithm for building the collection. This algorithm also assigns the unique identifiers for feeds, homepages and documents. Note that URLs had to be normalised in the same fashion as that performed at the crawling stage, and that any redirects experienced by the crawler while fetching that URL had to also be applied so the correct document was found. The permalinks crawler experienced 900,000 redirects while fetching permalink documents for the collection.

Additionally, the collection contains an extras directory, which contains files of useful information, including DOCNO to URL mapping tables, DOCNO to Date mapping tables, and a list of all URL to URL redirects the crawler experienced.

4. ANALYSIS OF COLLECTION

In this section, we show the statistics of the Blogs06 collection, as presented in Table 1. Over the eleven week time span of the feeds crawl, we collected over 753,000 RSS or Atom feeds, in which we found over 3.2 million perma-

Quantity	Value
Number of Unique Blogs	100,649
RSS	62%
Atom	38%
First Feed Crawl	06/12/2005
Last Feed Crawl	21/02/2006
Number of Feeds Fetches	753,681
Number of Permalinks	3,215,171
Number of Homepages	324,880
Total Compressed Size	25GB
Total Uncompressed Size	148GB
Feeds (Uncompressed)	38.6GB
Permalinks (Uncompressed)	88.8GB
Homepages (Uncompressed)	20.8GB

Table 1: The main statistics of the Blogs06 collection.

links. The total size of the collection is 148GB, consisting of 38.6GB of feeds, 88.8GB of permalink documents and 20.8GB of homepages. This compresses to 25GB.

Figure 2 shows the breakdown of feeds and documents into major components. Blog feeds from Blogspot and Livejournal compose the most substantial parts of the collection, and this is mirrored in terms of the number of documents fetched. In contrast, Xanga.com amounts to a noticeable percentage of feeds, yet produces few permalink documents.

4.1 Dates and Times

Figure 3 shows the number of URLs of permalink documents found each day. As would be expected, the number of documents found during each day of the first week of the feeds crawl is high, as all feeds were new, and hence all permalinks URLs found were new and had to be downloaded. After each feed had been visited once, the number of new documents settled down to a more regular pattern. Also noticeable on Figure 3 is the day on which there was a crawler failure, and seldom new permalink documents were found. When crawler failures happened, we tried to 'catch-up' on uncrawled feeds on following days.

Of the 3.2 million permalink documents in the collection, 136,344 permalinks did not have date information in the corresponding entry of their feed. A further 8,595 permalinks had unlikely dates before 1998, which is when the first XML feed started to appear on the Web. (Most of these had dates around 1970, which is the default time for most Unix/Linux computers when the time has yet to be set.) 279 permalinks had dates in 2007 and beyond. Table 2 shows the distribution of reported post dates across years. Hence, it appears that not all the dates reported in XML feeds can be trusted. Moreover, it would appear that we have a significant number of documents from outwith the time period of the crawl.

We desire to examine permalink documents that we know have definitely been posted in the time span of the crawl. However, because of the inaccuracies with the reported dates described above, we select documents that we are sure have been posted within the period of the collection: These are selected as documents with valid dates, in the time span of the collection, which was not more than one day ahead of the date at which the permalink URL was first identified in a feed, and not before the previous time that feed had been checked. Documents without dates in these ranges are

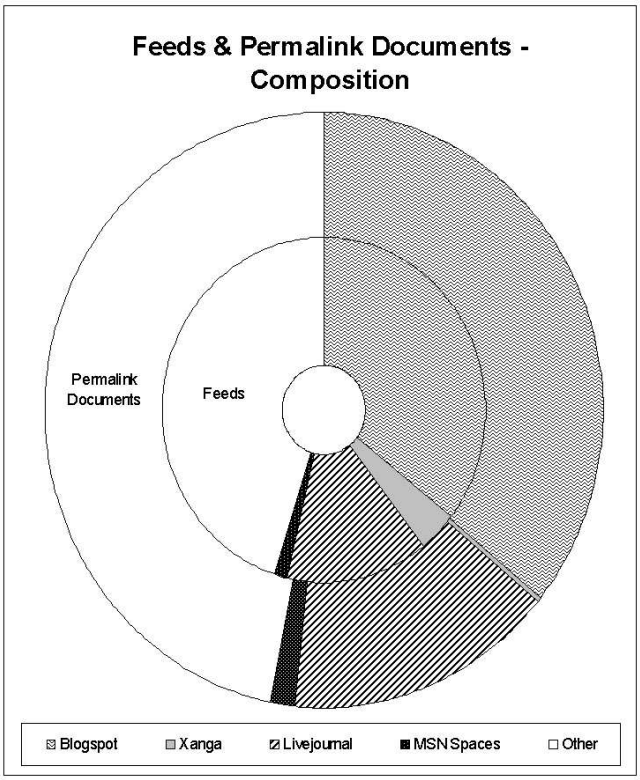


Figure 2: Doughnut plot showing the composition of the (inner ring) blogs selected to be in the Blogs06 collection, while the outer ring shows the composition of the permalink documents found during the time span of the collection. Blogspot and Livejournal make up the largest components of the collection. Xanga was a noticeably large component of the feeds, however comparatively few documents were found.

Year	Posts
<= 1980	8,591
1981-1990	2
1991-2000	95
2001	106
2002	494
2003	3,868
2004	32,717
2005	1,771,516
2006	1,261,159
2007-2010	108
2011-2020	88
2021-2030	65
2031 >=	18
(Missing)	(136,344)
(Total)	(3,215,171)

Table 2: Distribution of years from reported permalink dates

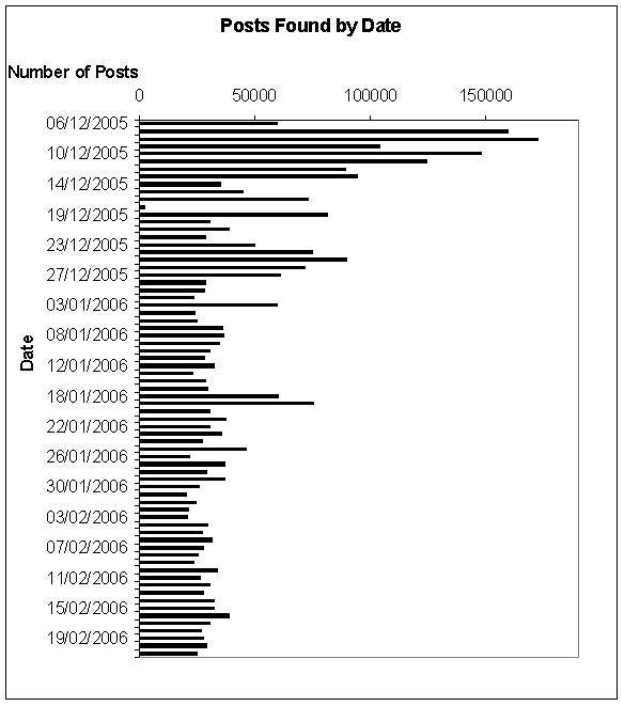


Figure 3: The number of documents found each day of the collection.

considered to have invalid dates or to have been created before the time period of the collection. From the 3.2 million permalink documents in the collection, only 1,929,647 documents have permalink documents with dates within the period of the collection (60%).

Figure 4 shows the number of documents for each day, from the reported date in the XML feed. We only use the 1.9 million documents identified above. Comparing this to Figure 3 above, we see a far more regular pattern, which is very similar to that observed in the BlogPulse collection by Kolari et al [8]. There is a noticeable weekly cycle, with far less blog posts made at the weekends. In addition, there is a pronounced dip in the number of postings for the December festive week - it appears that many bloggers appear to take a holiday from blogging over Christmas!

Finally, Figure 5 shows the distribution of postings (from the 1.9 million documents identified above), by reported hour of posting. All times are normalised to the local time of the poster. The figure shows that the highest volume of postings is around mid-afternoon, at 2pm, while the lowest volume is at around 3am. There is a distinct fall in hourly posting volume from midnight to 3am (88,000 posts to 52,000 posts), when it would appear that many bloggers go to bed. Overall, Figure 5 shows a similar time distribution to that observed for the Italian pings by Kolari et al in [8]. However, in our collection, it would appear that more bloggers blog later, as the fall off in posting rate in the evening is not as marked as that observed by Kolari et al.

4.2 Splogs

We originally inserted 17,969 known splogs into the Blogs06 collection (corresponding to 17.8% of the feeds). These splogs produced 509,137 documents (15.8%). For the re-

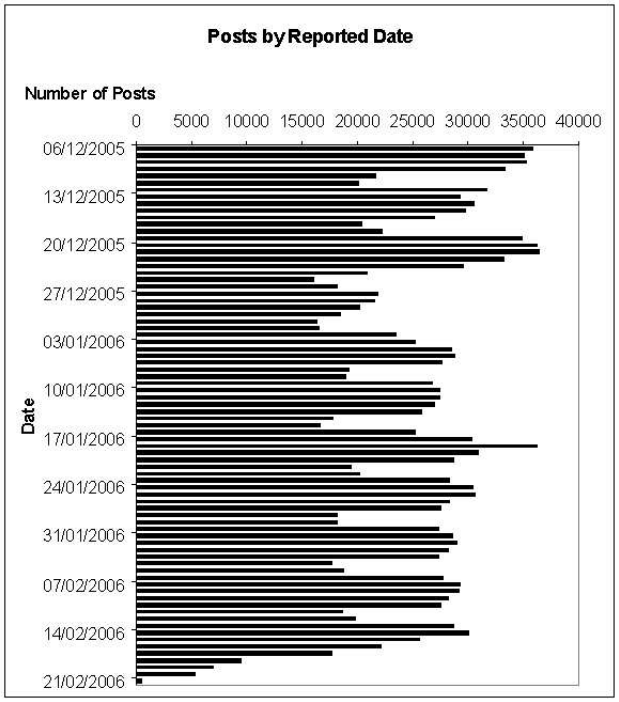


Figure 4: The number of documents posted by each day of the collection, from the date reported in the XML feed. This figure shows a cyclic distribution over week long period, with far fewer posts being made at the weekend. Note also the dip in postings in the 3rd week of the collection, during the festive Christmas holidays.

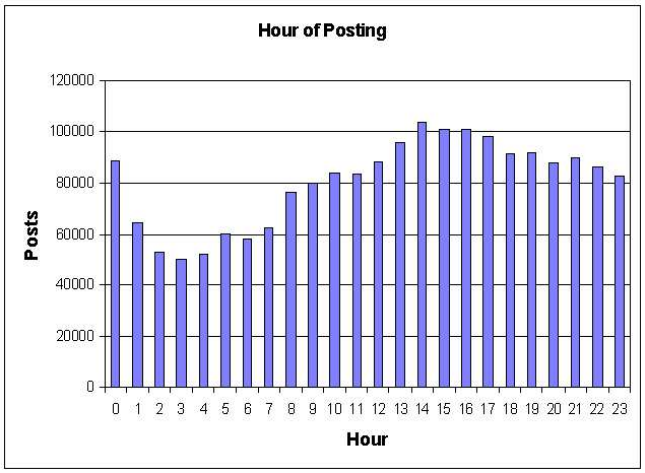


Figure 5: Posts made per hour over the time span of the Blogs06 collection.

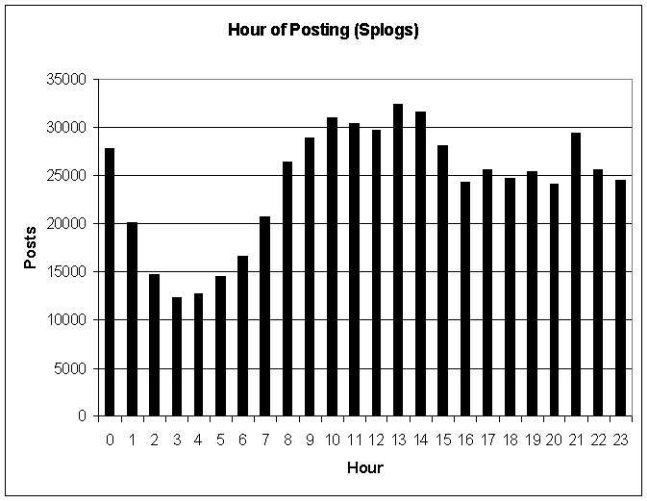


Figure 6: Post made per hour by known spam blogs in the Blogs06 collection.

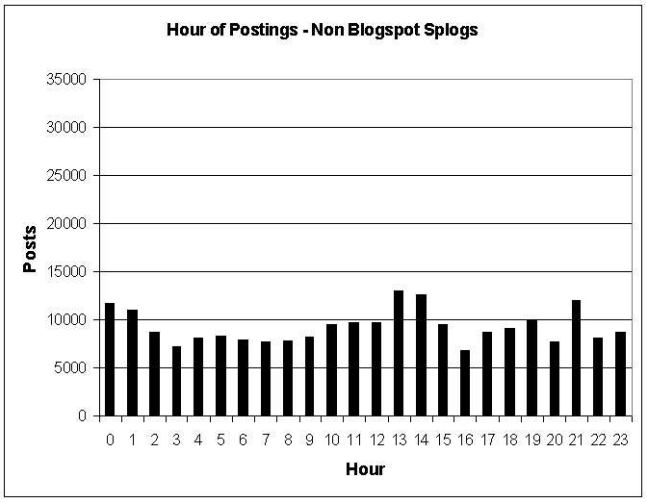


Figure 7: Posts made per hour by known spam blogs not hosted at Blogspot.com

mainder of this paper, we intend to study different characteristics of the Blogs06 collection, and assess how these characteristics vary across the splog and non-splog components of the collection.

Figure 6 shows the time distribution of postings by hour, from documents with valid permalink dates, but only from splogs. Comparing this to Figure 5, we can see that the time distribution of splog posts is similar to the distribution over all posts. Indeed, the Spearmans correlation coefficient between the distributions is 0.6643. Most splog posts seem to occur during the hours 8am to 3pm, suggesting that posts are only created semi-automatically, and such semi-automatic construction of splog posts occurs during there working hours. Our hypothesis here is that most splog sites in the Blogs06 collection are hosted at Blogspot.com (87%), which requires a Captcha ("*C*ompletely *A*utomated *P*ublic *T*uring test to tell *C*omputers and *H*umans *A*part") [16] to enter a post. Captchas are computer generated images of letters and numbers which users must correctly identify the contents of before being allowed to proceed with making the post. The Captcha images are distorted in such a way that computers find them extremely difficult to read, and this prevents spammers from making automatic splog posts. Instead, the spammer employs people to process captchas, or co-erces users of the spammer's own website in realtime to solve the Captchas, allowing blog posts to be submitted [17]. In such cases, the hour distribution of Blogspot posts would follow the same hour distribution as that of the employed person population.

To validate our hypothesis, we plot the hour distribution of splog posts that did not originate from Blogspot. This is shown in Figure 7. This shows, in contrast to Figure 6, no concernable pattern, meaning that the majority of the non-Blogspot splogs are not employing Captcha techniques. Figure 7 is similar to the time distribution of splog pings observed by Kolari et al in [8], in the lack of a discernible pattern. Indeed, the Spearmans correlation coefficient between the distributions of Blogspot splog posts and non-Blogspot splog posts is 0.3383, showing that there is little correlation in the rankings - the hour distribution of Blogspot splog posts is somewhat different to the non-Blogspot splog posts. Interestingly, the Captcha system described above was implemented by Blogspot between the time spans of the BlogPulse collection examined by Kolari et al and the Blogs06 collection described in this work.

5. PING COVERAGE

Instead of polling feeds once a week, an alternative method of constructing the collection would have involved the use of pings. Pings are XML-RPCs (XML Remote Procedure Calls) made from blog publishers to blog ping servers. The purpose of each ping is to inform the ping server (usually a blog search engine) that the blog's feed has been updated. Ping servers often make their ping log available in real time, to allow other search engines to receive notifications of blog updates.

We monitored the PubSub.com ping log while building the collection, from 06/12/2005 to 01/02/2006. This corresponds to 57 days of the 77 days of the collection. The ping log for this period has 40 million pings. Of the 100,649 feeds in the Blogs06 collection, 37% are represented in the collected ping log: 76% of the splogs are present in the ping log, while only 41% of the non-splogs are present. This shows that the PubSub.com ping log would have given much less coverage of the subset of the blogsphere selected to be included in the Blogs06 collection, when compared to polling each feed for updates.

The feeds present in both the Blogs06 collection and the ping log received a total of 541,071 pings, meaning that there should be the same number of documents for those feeds. In contrast, we detected 1,010,484 documents for the same blogs over the same time span as the ping log, showing that the ping server did not always receive a ping when those blogs were updated.

6. TERM FEATURES

Due to the presence of splogs in the Blogs06 collection hosting content of an adult nature, we examine the presence of offensive terms in the corpus. The hypothesis here is that

Offensive Words	Spam	Ham	Total
in Documents	12.2%	14.8%	14.4%
in URLs	1.1%	0.8%	0.8%

Table 3: Percentage of documents in the spam and non-spam (ham) documents of the collection containing offensive English terms.

nude	*free*	*teen*	*girl*
insur	*medic*	*decor*	*gambl*
bodi	*weight*	*school*	*credit*
casino	*adult*	*card*	*mortgag*
poker	*game*	*nake*	*health*
porn	*women*	*interest*	*gift*
discount	*offshor*	*rate*	*pregnanc*

Table 4: A selection of top informative terms from the spam documents of the collection. Informative terms were identified using term frequency divergence between the spam documents and the entire collection.

offensive terms would be more prevalent in the documents from splogs. Using an offensive words list supplied by a major British broadcaster, we examine the presence of offensive terms in both the URL and the content of the documents. The list of words used includes acceptable longer versions of the offensive terms. (i.e. the term Arsenal - a British football team - is acceptable, but the first four letters are not).

Table 3 shows the percentage of documents and URLs containing offensive terms, in terms of spam documents from known splog feeds, and the ham (non-spam) documents. The results are interesting in that both document groups have offensive content in similar amounts. This can be interpreted in two ways: that the splogs are trying to keep similar content terms to normal blogs to avoid detection. Moreover, that bloggers can feel strongly about a topic they are blogging about, that they feel they need to use offensive words to make their point.

Furthermore, by measuring the term frequency divergence between the spam documents and the collection, we examine the most informative terms in the spam documents. Table 4 lists some of the most informative terms from the spam documents. Note that these terms are stemmed forms, as per the Porter stemming algorithm.

7. LINK STRUCTURE

Finally, we examine the link structure within the Blogs06 collection. Blogs show a power-law of inlink and outlink distributions [8], which is typical of the Web in general [5]. Figure 8 shows the in-degree and out-degree distributions for all permalink documents in the Blogs06 collection. From this figure, we can see that the out-degree distribution exhibited in the collection is a smooth power-law distribution, while that of the inlinks is less smooth. Notice that some blog posts have over 10,000 out links in one page, possibly caused by unchecked comment spamming (comment spam not removed by the blogger). Comment spam is not a feature of the BlogPulse collection, as all posts were collected as soon as possible after BlogPulse was pinged to inform them that the blog had been updated.

By breaking the in-degree and out-degree distributions down into spam and non-spam (ham) permalink documents, we aim to see if the link distributions are noticeably different. Figures 9 & 10 show the in-degree and out-degree distributions for the spam and non-spam (ham) documents respectively. From Figure 10 it is clear that the in-degree and out-degree graphs of the collection are being dominated by the ham component of the collection. In contrast, the spam in-degree and out-degree distributions are markedly different, with upward trends at out-degree 1000. We hypothesise here that the splogs have themselves become victims of unchecked comment spamming.

Also of note are the spikes around in-degree 15 (approximately) in all figures. We hypothesise that this is caused by the archive of "Recent Posts" most blogging software creates, and by removing links between the same blog's documents from our link graphs, these spikes would be removed.

8. RELATED WORK

There have been several Web test collections created for the purposes of Web data evaluation at TREC. These range from VLC, VLC2 through to .GOV and .GOV2. In particular, Bailey et al [3] documented the building of the WT10G collection from the VLC2 crawl of the Internet.

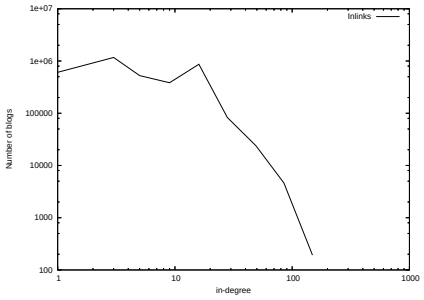
In 2005, BlogPulse released a test dataset of their blog data, spanning a 21 day period of July 2005. This contrasts from the Blogs06 test collection described in this paper in several ways: this collection is smaller in terms of time span; and as documents were collected immediately after BlogPulse was informed of their presence by pings, they generally do not include comments. This collection was studied by several participants for the 3rd Annual Workshop on the Weblogging Ecosystem [2]. For example, see [8].

Other studies of the Blogsphere have used their own private corpora of blog data. Nanno et al [13] supplemented a ping log with Web crawling to identify blogs. Of particular note is IBM's Webfountain project, which maintains large collections of online material, including postings from 300K blogs [7, 6].

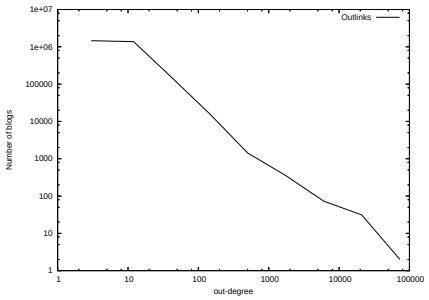
9. CONCLUSIONS

In this paper, we described the motivations, properties and construction of the Blogs06 collection for the TREC Blog track. Furthermore, we examined the statistics of the collected data, in particular the dates and times of posts. We examined how the distribution of dates and times altered between the known spam documents and the ham documents. Moreover, we analysed the coverage of the collection by the PubSub.com ping log. Next, we looked at the occurrences of offensive terms in the spam and ham components of the collection, and derived a list of top terms from spam blog posts. Finally, we considered the link structure in the Blogs06 collection, and showed that while the link structure in the spam component did not form a power law distribution, the non-spam documents did.

The Blogs06 test collection is a wide-ranging collection suitable for many research tasks relating to the blogsphere. This paper described the building of the collection, and provided an overview of the properties of the collection, that are of interest in the research community.

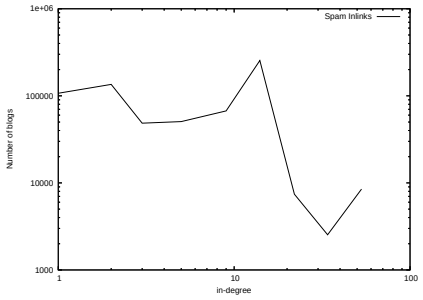


(a) Inlink Distribution over all Permalink Documents

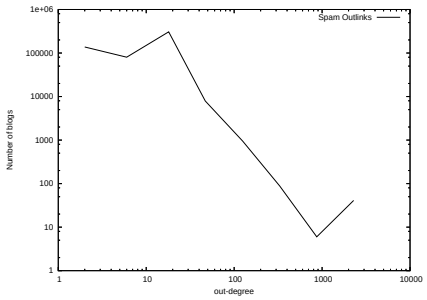


(b) Outlink Distribution over all Permalink Documents

Figure 8

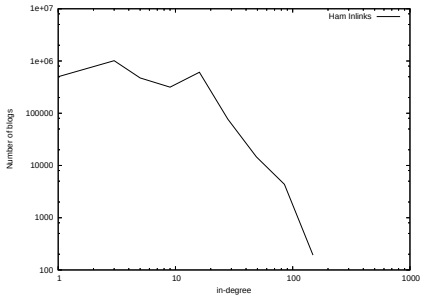


(a) Inlink Distribution over Spam Permalink Documents

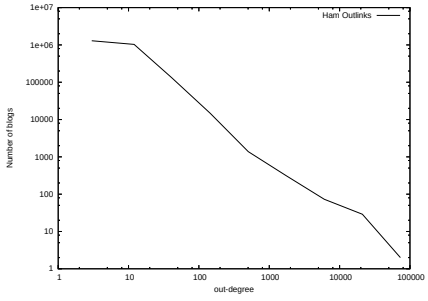


(b) Outlink Distribution over Spam Permalink Documents

Figure 9



(a) Inlink Distribution over non-Spam Permalink Documents



(b) Outlink Distribution over non-Spam Permalink Documents

Figure 10

10. REFERENCES

[1] HTML author's guide to the robots meta tag. `www.robotstxt.org/wc/meta-user.html`, last accessed 11/05/2006.

[2] E. Adar, N. Glance, and M. Hurst, editors. *Proc. of 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis & Dynamics*, May 2006.

[3] P. Bailey, N. Craswell, and D. Hawking. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing and Management*, 39(6):853–871, 2003.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[6] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *SIGKDD '05:*, 78–87, 2005.

[7] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. WWW '03: World Wide Web*, 491–501, 2004.

[8] P. Kolari, A. Java, and T. Finin. Characterizing the Splogosphere. In [2].

[9] M. Koster. A standard for robot exclusion. `www.robotstxt.org/wc/norobots.html`, last accessed 11/05/2006.

[10] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. WWW '03: World Wide Web*, 568–576, 2003.

[11] M. Meeker and B. Pitz. An Update from the Digital World - October 2004 `www.morganstanley.com/institutional/techresearch/pdfs/dw_syndication1004.pdf`, last accessed 11/05/2006.

[12] G. Mishne and M. de Rijke. A study of blog search. In *Proc. ECIR 2006*, 289–301. April 2006.

[13] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura. Automatically collecting, monitoring, and mining japanese weblogs. In *Proc. WWW Alt. '04: World Wide Web posters*, 320–321, 2004

[14] A. Rosenbloom. The blogosphere: Introduction. *Commun. ACM*, 47(12):30–33, 2004.

[15] D. Sifry. State of the Blogosphere, April 2006 Part 1: On Blogosphere Growth. `technorati.com/weblog/2006/04/96.html`, last accessed 11/05/2006.

[16] L. von Ahn, M. Blum, and J. Langford. Telling humans and computers apart automatically. *Commun. ACM*, 47(2):56–60, 2004.

[17] N. R. Wagner. Capthas and information hiding. Technical report, University of Texas at San Antonio, 2003. `www.cs.utsa.edu/~wagner/captcha/hiding12.pdf`, last accessed 10/05/2006.