Everything you would like to know about RSS feeds and you are afraid to ask

Zeinab Hmedeh, Nicolas Travers, Nelly Vouzoukidou, Vassilis Christophides, Cédric Du Mouza, Michel Scholl

Everything you would like to know about RSS feeds and you afraid to ask

Zeinab Hmedeh[1], Nicolas Travers[1], Nelly Vouzoukidou[2], Vassilis Christophides[2], Cedric du Mouza[1], and Michel Scholl[1]

[1] CEDRIC Laboratory - CNAM - Paris, France
firstname.lastname@cnam.fr
[2] FORTH/ICS and Univ. of Crete - Heraklion, Greece
(christop,vuzukid)@ics.forth.gr

Abstract. We are witnessing a widespread of web syndication technologies such as RSS or Atom for a timely delivery of frequently updated Web content. Almost every personal weblog, news portal, or discussion forum employs nowadays RSS/Atom feeds for enhancing traditional *pull-oriented* searching and browsing of web pages with *push-oriented* protocols of web content. Social media applications such as Twitter or Facebook also employ RSS for notifying users about the newly available posts of their preferred friends (or followees). Unfortunately, previous works on RSS/Atom statistical characteristics do not provide a precise and updated characterization of feeds' behavior and content, characterization which can be used to successfully benchmark effectiveness and efficiency of various RSS/Atom processing/analysis techniques. In this paper, we present the first thorough analysis of three complementary features of real-scale RSS/Atom feeds, namely, publication activity, items structure and length, as well as, vocabulary of the textual content which we believe are crucial for Web 2.0 applications.

Keywords: RSS/Atom Feeds, Publication activity, Items structure and length, textual vocabulary composition and evolution

1 Introduction

Web 2.0 technologies have transformed the Web from a publishing-only environment into a vibrant information place where yesterday's end users become nowadays content generators themselves. Web syndication formats such as RSS[3] or Atom[4] emerge as a popular mean for timely delivery of frequently updated Web content. According to these formats, information publishers provide brief summaries of the content they deliver on the Web, called *items*, while information consumers subscribe to a number of RSS/Atom *feeds* (i.e., streams or channels) and get informed about newly published items. Today, almost every

[3] web.resource.org/rss/1.0/
[4] tools.ietf.org/html/rfc5023

personal weblog, news portal, or discussion forum employs RSS/Atom feeds for enhancing traditional *pull-oriented* searching and browsing of web pages with *push-oriented* protocols of web content. Furthermore, social media applications such as Twitter or Facebook also employ RSS for notifying users about the newly available posts of their preferred friends (or followees).

Unfortunately, previous works on RSS/Atom statistical characteristics [15, 27, 13] do not provide a precise and updated characterization of feeds' behavior and content which could be effectively used for tuning refreshing policies of RSS aggregators [24, 22], benchmarking scalability and performance of RSS continuous monitoring mechanisms [19, 9, 7, 8, 25, 5] or comparing various techniques for RSS items mining, recommendation, enrichment and archiving [3, 26, 28]. In this paper, we present the first thorough analysis of three complementary features of real-scale RSS/Atom feeds, namely, *publication activity*, *items structure and length*, as well as, *vocabulary of the textual content*.

Our empirical study relies on a large-scale testbed acquired over a 8 month campaign from March 2010 in the context of the French ANR Roses project[5]. We collected a total number of 10,794,285 items originating from 8,155 productive feeds (spanning over 2,930 different hosting sites) which extraction is further explained in section 2.

The main conclusions drawn from our experiments are:

1. As analyzed in Section 3, 17% of RSS/Atom feeds produce 97% of the items of the testbed. In their majority, productive feeds (*i.e.* with >10 items per day) exhibit a regular behavior without publication bursts, thus are more predictable in their publication behavior. As expected, micro-blogging feeds from social media are more productive than those from blogs while press sources lie in between. The average publication rate among all feeds of the testbed has been measured to be 3.59 items a day;
2. As highlighted in Section 4, the most popular RSS/Atom textual elements are *title* and *description* while the average length of items is 52 terms (which has not been reported so far in related work). It is clearly greater than bids (4-5 terms [10]) or tweets (∼15 terms at most [21]) but smaller than blog posts (250-300 terms [16]) or Web pages (450-500 terms excluding tags [14]). In addition, re-publication of items across feeds is rare since we identified only 0.41% of duplicates among distinct feeds hosted by different sites;
3. As studied in Section 5, the total number of extracted English terms is 1,537,730 out of which only a small fraction (around 4%) is found in the `WordNet` dictionary. This is due to the heavy use in RSS/Atom textual elements of named-entities (person and place names), URLs and email addresses as well as to the presence of numerous mispellings, spoken acronyms or special-purpose jargon which are phenomena related to the informal and ephemeral use of language in the Web 2.0 era. We formally characterized the total vocabulary growth using Heap laws as well as the number of occurrences of the corresponding terms with a *stretched exponential distribution* used for the first time in the literature in this respect. We observed that the

[5] www-bd.lip6.fr/roses

ranking of vocabulary terms does not significantly vary during the 8 month period of the study for frequent terms.

2 Feed acquisition and Warehousing

RSS standard aims at exchanging on a day-to-day basis concise summaries of the information published on the Web such as news headlines, search results, "What's New", job vacancies, and so forth. A large amount of this frequently updated content can be thought of as a list which users need to keep track of and customize. RSS has an XML-based format that allows the *syndication* of lists of hyperlinks, that helps viewers to decide whether they want to follow a link. To enable this functionality, a Web site needs keeping a `feed`, or `channel`, available, just like any other file or resource on the server. A feed contains a list of `items` or `entries`, each of which being identified by a URL. Items typically are summaries of distinct information pieces to be found elsewhere on the Web and can have any amount of other *metadata* as well. Once a feed is available, web applications can regularly fetch the file to get the most recent items on the list. In essence, an RSS feed is an URL that returns an XML document in an accepted RSS format, possibly with informal additions, containing at its heart a list of 'items' carrying its main content. Feeds can also be used for other kinds of list-oriented information, such as syndicating the content itself (often weblogs) along with the links. Atom is an effort of IETF to come up with a well-documented, standard syndication format. Although it has a different name, it has the same basic functions as RSS, and many people use the term "RSS" to refer to RSS or Atom syndication.

For our study we have extracted about 100,000 feeds from major RSS/Atom directories, portals and search engines (such as syndic8.com, completeRSS.com, Google Reader, feedmil.com, retronimo.com or thoora.com) over an eight month campaign from March 2010. Out of these feeds, only 12,611 were exempt of communication errors and could be successfully validated against RSS/Atom formats. The major cleaning decisions made were related to the facts that (a) a huge part of feed URLs did not exist any more or were modified, (b) some of them were accessible only through user sessions, (c) numerous feeds had missing or ill-formed XML tags or miss-encoding files. From these 12,611 resulting feeds, 27,003,115 items were warehoused into a MySQL local database[6] on which we have extracted non-intra-replicates feeds with 10.7 million items. Since item date, terms and category are crucial for the following analysis and feed characterization, a particular attention was paid to the quality of the items along the three following lines: For our studies, we paid attention on the quality of our testbed:

- In order to characterize the composition of the testbed in terms of information sources of feeds (Social Media, Press, Forum, Sales, Misc. and Blog -

[6] Available on line at deptmedia.cnam.fr/∼traversn/roses

further details in section 3) we rely on a manual classification of feeds based on the available information from their metadata (title, category, link or description tags) by recognizing special keywords (*e.g.* currencies, forum, daily news), specific hosting domains/url (*e.g.* twitter, blogspot, yahoo), special items composition (*i.e.,* a single pattern for all items). Automated extraction of summaries of feeds like in [3] is beyond the scope of our study;
- We guarantee that for each feed only one item occurrence is kept. To this end we rely on exact matching semantics between items using a hash function on the content of title and description tags. During this tasks we have isolated some highly redundant feeds that essentially do not respect the chronological order assumption by re-delivering previous feed items (*i.e.* random or hot topics), or that change their content automatically according to a pattern (*i.e.* number of users' views on the items, item's current rating);
- Since, the publication date (`pubDate`) is missing in 20% of items, we have associated a timestamp to each item. It is computed as the time elapsed from the insertion of the very first item into the warehouse. This ensures a realistic reproduction of the original items flow.

3 Feeds Analysis

In this Section we are interested in characterizing the composition of our testbed in terms of the type of the information source RSS/Atom feeds are originating from, as well as, in studying their corresponding publication activity. Although global statistics regarding Web 2.0 activity are constantly monitored[7], a in depth analysis of RSS/Atom feeds productivity in relation to their source type has not already reported in the literature. Knowing that highly active feeds are more predictable in their publication behavior, would guide for instance resource allocation mechanisms to be tied to the feeds category.

3.1 Source Type

Six types of information sources delivering their content as RSS/Atom feeds were identified: `Press` (for newspapers and agencies), `Blogs` (for personal weblogs), `Forums` (for discussion and mailing lists), `Sales` (for marketing web sites), `Social media` (for micro-blogging such as twitter, digg, yahoo! groups, and blogosphere) and `Misc.` (*e.g.*, for news sites with medical or city/group information as well as podcasts). Then, the 8,155 productive feeds were manually classified under these six types as described in Section 2.

Table 1 depicts for each type, the corresponding percentage of feeds and items as well as the average number of items per feed. `Social media`, `Press` and `Forums` are more productive (with an average number of items per feed ranging from 3178.01 to 7085.03) than `Sales`, `Misc` and `Blogs` (with less than

[7] thefuturebuzz.com/2009/01/12/social-media-web-20-internet-numbers-stats

Type	% of feeds	% of items	$\frac{\#\ items}{\#\ feeds}$
Social Media	1.77%	9.45%	7085.03
Press	9.99%	38.82%	5141.24
Forum	1.51%	3.62 %	3178.01
Sales	11.32%	15.49%	1811.92
Misc.	41.47%	25.47%	812.99
Blog	33.93%	7.14%	278.55

Table 1. Source types of RSS/Atom feeds

2000 items on average per feed). As discussed in the following sections, feeds' behavior can be further refined by considering daily publication rates as well as the corresponding activity variability over time.

The issue of this empirical work seems representative of the Web reality. In fact, with a predominance of blogs feeds opposed to social medias which are very productive feeds compared to the number of feeds, and the high number of press and sales feeds makes sense in our testbed. A sight of the Web 2.0 productivity has been reported further[8] but no other empirical works have provided such kind of characterization over a wide classification of feeds.

3.2 Publication Activity

A time-agnostic characterization of RSS feeds activity has been originally proposed in [24] by studying the distribution of the number of feeds publishing at a given rate x. A similar power law behavior ax^b was observed for the 8,155 feeds, although with different coefficients $a = 1.8 \times 10^3$ and $b = -1.1$. This is due to the presence in our testbed of more productive feeds than in [24] (featuring more blog-originating feeds). This postings activity is depicted in Figure 1, in which we can see the power law distribution that enlights the fact that most feeds produce few items. By the way, this power law do not take into account the long tail of the distribution corresponding to very productive feeds which will represent most of the produced items in the testbed.

A coarse-grained characterization of the temporal variation of feeds activity has been suggested in [27] which analyzes publication burstiness. With a similar burst definition of at least 5 times the average publication rate during a unit of time (a day), 89% of our feeds produce bursts. However, the remaining 11% of feeds without bursts produce more than 81% of the items. Thus, this burstiness measure is not sufficient for a precise characterization of feeds activity, especially for very productive feeds.

We prefer the *Gini coefficient* [20], denoted as G, to characterize the variability of feeds' publication activity over time. G is adequate for time series (feeds are time series), since it does not only take into account the deviation from a

[8] Web 2.0 Stats: http://thefuturebuzz.com/2009/01/12/social-media-web-20-internet-numbers-stats/

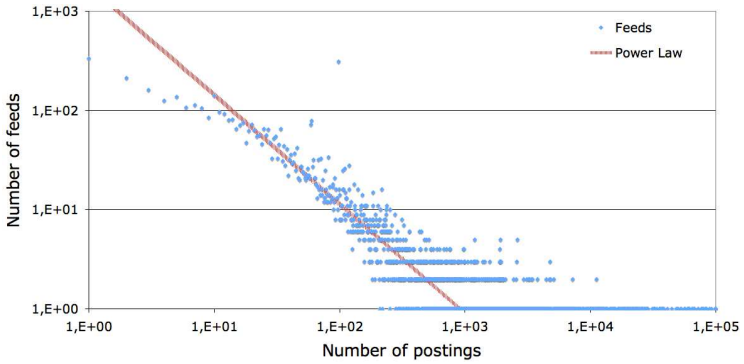

Fig. 1. Number of Postings vs Number of correspondings feeds

mean value (as in the case of burstiness) but also the temporal variation of this deviation. It has been widely used for analyzing a variety of phenomena arising in economy, geography, engineering, or in computer science (e.g., self-join size, supervised learning). G is defined as:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j|}{2 \times n \times \sum_{i=1}^{n} y_i}$$

with y_i denotes the number of items, sorted in increasing order, over the n days. A G value close to 1 suggests that the number of items of a feed significantly varies over time.

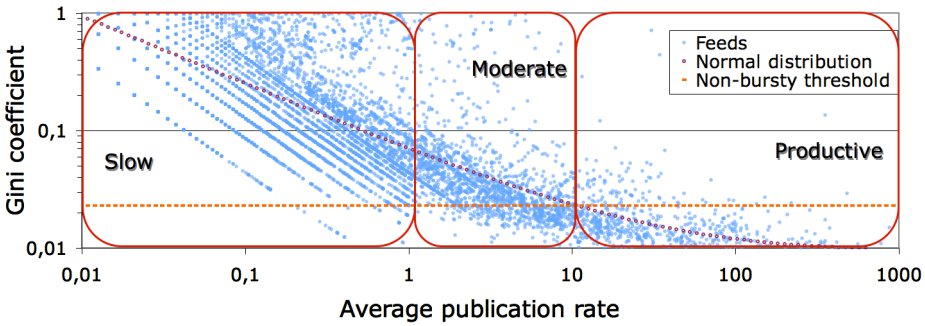

Fig. 2. Activity classes of RSS/Atom feeds

Figure 2 depicts in log-log-scale G vs the average publication rate for each of the 8,155 feeds. Feeds with a G value less than 0.02293 (below the horizontal dashed line) did not exhibit a single burst during the entire 8-month period. Three classes of feed activity are identified. The first one called "`productive class`", comprises 614 feeds which produce more than 10 items a day with a

very low temporal variation (G less than 0.03). The second one called "`moderate class`", gathers 1,677 feeds that publish between 1 and 10 items per day with a moderately low temporal variation (G less than 0.1). The third one called "`slow class`", represents the large majority of the feeds. They publish less than 1 item a day (0.23 on average) and exhibit a strong temporal variation in the number of items ($G = 0.32$ on average). The average G value as a function of the average publication rate p (dotted curve) is approximated by the following function:

$$G(p) = 5.53 \times 10^{-2} \times p^{-0.59} + 4.48 \times 10^{-3} \times p^{0.101}$$

Each class characteristics are detailed in Table 2 which confirms this sight. Productive feeds are variable but according to its high publication rate the Gini coefficient is low. At the opposite, slow feeds have a low variance and a higher average Gini coefficient showing that they are extremly bursty feeds.

	`productive class`	`moderate class`	`slow class`
avg. pubRate	61.816	3.592	0.308
variance	14,241.24	5.72	0.062
covariance	4.71×10^{-8}	3.96×10^{-9}	5.47×10^{-10}
max. nb of items	11,314	1,235	64
avg *Gini* coeff	0.0204	0.0596	0.1321

Table 2. Daily characteristics of the different activity classes

In order to illustrate the tested, Table 3 gives a sample of urls from each class. We show here some representative feeds from each class according to its type, average publication rate and Gini coefficient.

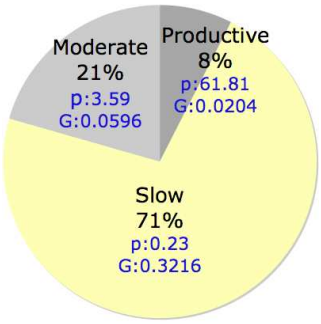

Fig. 3. Feeds per activity class

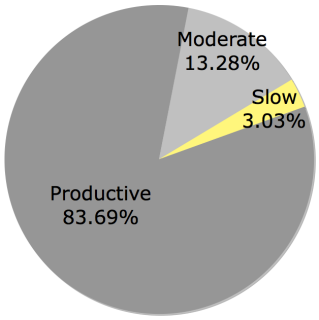

Fig. 4. Items per activity class

Unlike [15] reporting that 57% of the feeds have a publication rate > 24 items a day, in Figure 3 we can see that only 8% of our 8,155 feeds were actually very productive (with > 10 items a day) while 71% had a low publication activity.

Activity Class	URL	type	rate	coeff
productive	rss.sportsblogs.org/rss/daily.xml	Social	1275	0.008
	feeds.feedburner.com/Techcrunch	Social	25	0.012
	press.jrc.it/rss?id=all.rss&type=bns	Press	1184	0.009
	preciseNews.us/pt/planetrss?group=news	Press	471	0.009
	feeds.feedburner.com/fixya/wgvY	Misc	593	0.008
moderate	publisherdatabase.com/forums/rss.php?t=1	Forum	2.26	0.029
	csmonitor.com/rss/usa.rss	Press	9.6	0.017
	lemonde.fr/rss/sequence/0,2-3234,1-0,0.xml	Press	9.4	0.027
	blog.reidreport.com/feed	Blog	5.31	0.016
	medicalnewstoday.com/rss/cancer-oncology.xml	Misc	8.88	0.021
slow	tugsearch.co.uk/blog/feed	Blog	0.90	0.048
	koantum.wordpress.com/feed	Blog	0.10	0.993
	download-soft.com/rss-feeds/new-Audio.xml	Sales	0.96	0.597
	fashionjewelryforeveryone.com/Under5ERRSS.xml	Sales	0.86	0.993
	feeds.ezinearticles.com/expert/Smit_Chacha.xml	Misc	0.88	0.114

Table 3. Extracted URLs from activity classes

Note also that from the initially harvested 12,611 feeds around 35% did not publish any item during the entire 8-month period. Of course productive feeds, even though they represent a minority, account for the majority of the items harvested. We see in Figure 4 that feeds of `productive class` publish 83.69% of the total number of items while the majority of feeds from `slow class` produce only 3.03% of the items (Power Law behavior of feeds' publication rate).

Type	productive class			moderate class			slow class		
	%	p	G	%	p	G	%	p	G
Social Media	**44.9%**	65.33	.015	8.1%	5.57	.026	**47.0%**	0.18	.470
Press	27.4%	70.76	.020	**43.0%**	3.61	.036	29.6%	0.34	.237
Forum	21.0%	54.51	.018	20.2%	3.85	.042	**58.8%**	0.29	.348
Sales	8.3%	83.64	.022	19.0%	3.43	.122	**72.7%**	0.23	.454
Misc	2.7%	63.12	.024	12.9%	3.70	.059	**84.4%**	0.23	.338
Blog	4.3%	15.53	.019	24.7%	3.37	.056	**71%**	0.22	.267

Table 4. Source types and activity classes of feeds

Last but not least, Table 4 provides for each source type and activity class, the corresponding percentage of feeds, the average publication rate (p), as well as, the average Gini coefficient (G). Feeds from Social Media are almost evenly distributed in `productive` and `slow` classes with a temporal variation in their publication rate which is more important in the latter than in the former class. This is related to *Twitter* like behavior with *Followers* and *Followees* (users can follow - slow class - or users can be followed and produce a lot - productive class). Press feeds exhibit a `moderate` and almost *regular* publication activity,

while feeds originating from Forums, Sales, Misc and Blogs sites mainly exhibit a `slow` publication activity. It is worth also noticing that in this class, Sales are mostly bursty feeds while Blogs and Forums exhibit a more regular behavior.

4 Items Analysis

This Section successively focuses on the analysis of items' structure and length as well as on the items' replication rate across RSS/Atom feeds.

4.1 Item Structure

The empirical analysis presented in Table 5, reveals that a great number of fields (tags) foreseen in the XML specification of RSS or Atom formats are actually not utilized in items. While `title`, `description` and `link` are present in almost all the items, `pubDate` is missing in around 20% of the items, and the `language` information is missing in 30% of the feeds. Almost 2/3 of the items are not *categorized* while `author` information is present in less than 8% of the items. It is worth noticing that 16% of fields contain errors or are used with a wrong format (mixing RSS or Atom fields). Other XML tags are only sparsely used in our testbed and are not reported here. In a nutshell, RSS/Atom items are characterized by the predominance of textual information as provided by `title` and `description` over factual information provided by `category` and `author`.

title	**link**	**pubDate**	**desc.**	**Language**
99.82%	99.88%	80.01%	98.09%	69.14% (feed)
author	**category**	**GUID**	**ext.**	**RSS/Atom**
7.51%	33.94%	69.50%	29.73%	16.48%

Table 5. Popularity of XML tags in RSS items

4.2 Item Length

Next, we focus on the length of items measured as the number of terms of the textual `title` and `description` fields. Items are short with a 52-terms size on the average (see Table 6). The high variance (11,885.90) of items length is mainly due to the large diversity of the description fields that can be either missing (or be a simple *url*) or oppositely be a long text (even an entire HTML document).

Figure 5 plots the number of items vs their length. A long-tail curve is observed in items length distribution as also reported in the literature for the size of Web documents [30]. 51.39% of the items have a length between 21 and 50 terms, and 14% between 8 and 20 terms. The peaks for length 6 and 8 are mainly due to `Sales` feeds (producing >55% of the items for these lengths) whose items respect a fixed-length pattern (*i.e.* items differ only by one or two terms).

	title + description	title	description
Average	52.37	6.81	45.56
Max	10,262	235	10,256
Variance	11,885.90	12.97	11,821.36

Table 6. Textual content characteristics in items

The main conclusion from this analysis is that RSS/Atom items are longer than average advertisement bids (4-5 terms [10]) or tweets (∼15 terms at most [21]) but smaller than the original blog posts (250-300 terms [16]) or Web pages (450-500 terms excluding tags [14]).

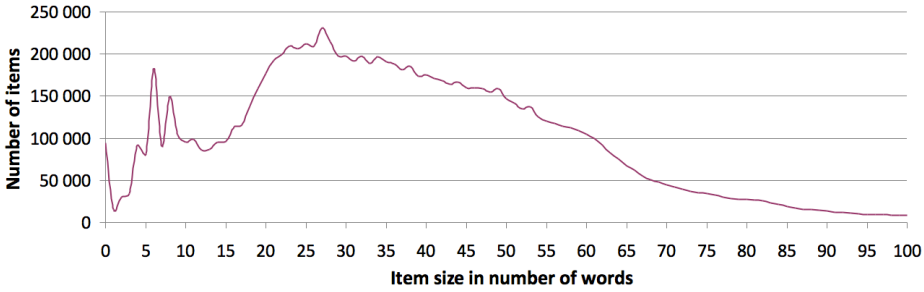

Fig. 5. Number of items per item length

4.3 Item Replication

Finally, we stick our attention to the replication of items either by feeds hosted by the same site (*intra-host* replication) or across different feeds (*inter-feed* replication). Replication detection is performed by exact matching of the item content based on a hash function. Out of the originally harvested 27 millions of items, 10.7 million distinct items per feed were extracted. Eliminating *intra-host* and *inter-feed* replication makes this number drop to 9,771,524 of items. Correlated to this, we identified that replicated items are naturally mostly present in the `productive class` with 85% of them, and 10% for the `moderate class`.

% of replication	< 10%	10 - 19%	20 - 29%	≥30%
% of hosts	95.19%	1.09%	0.68%	3.04%
% of items	71.31%	10.88%	10.23%	7.58%

Table 7. % Hosts/items per intra-host replication

Since some dedicated websites replicate information in several feeds, we want to show this behavior. For this, we have to study intra-host (same IP address) replication by identifying exact match between items produced by a same host. We gathered every sub-hosts under the corresponding single hosts (*i.e.* "rss.nytimes.com" and "blogs.nytimes.com" become "nytimes.com"). Table 7 reports the importance of *intra-host* replication that was measured over the 8 months period for the 2,930 hosts (with distinct IPs) of the 8,155 feeds. 95% of the hosts (which publish 70% of the items) have less than 10% of replication. Only a few hosts (less than 5%) of Sales, Press (especially news agencies) and Blogs feeds publish most of the replicated items (replication rate > 30%).

# distinct feeds	Once	Twice	$\geq$ 3 times
% items	99.51%	0.41%	0.08%

Table 8. % of items replicated across feeds

Once intra-host replicates had been removed, *inter-feed* item replication was measured. Table 8 reports for each replicated item the number of *distinct* feeds in which it appears. Clearly replication across feeds in different hosts is negligible: it accounts for less than 0.5% of the total number of items. This behavior can be explained by the absence in the testbed of RSS aggregators such as GoogleReader[9] or Yahoo Pipes[10] which systematically replicate other feeds.

Type	**Once**	**Twice**	**3 times**	**> 4 times**
Social	89,34%	7,76%	2,87%	0,03%
Press	83,94%	10,62%	3,67%	1,77%
Forum	88,51%	11,47%	0,02%	0,00%
Sales	86,42%	9,64%	1,46%	2,48%
Blog	88,79%	8,01%	2,41%	0,79%
Misc	79,55%	10,65%	7,29%	2,52%

Table 9. Intra-type replication: percentage of items replicated X times

To finish with, lets focus on the ratio of replication per item within each type. Table 9 shows, for each type, the percentage of items replicated a given number of times (once, twice, etc...). Mostly, items are never replicated but rates of multiple replication varies from a type to another. We can notice that, Misc, Forum and Press have the highest rates of double replication while Sales, Misc and Press keep a non-negligible rate of multiple-replication rate.

[9] www.google.com/reader
[10] pipes.yahoo.com

5 Vocabulary Analysis

This Section focuses on the analysis of the vocabulary of terms extracted from the title and description fields of RSS/Atom items. In order to deduce information regarding the quality of the employed vocabulary (valid terms/typos), we restrict our analysis to the English language, since a large majority of the analysed items in our testbed was written in English. In addition, as previous studies report, a similar behavior to that of the English language is exhibited for text corpora written in different languages (e.g. French, Korean [4], Greek [6], etc.). An automatic filtering of feeds (items) written in English turned out to be a difficult task given that language information was not always available.

7,691,008 "English" items were extracted as follows: an item is considered as English if the language tag exists in the feed metadata and is set to "en", or the feed url belongs to an English spoken host (*e.g.* "us" or "uk"), or a ".com" feed without any language tag. Then, a lexical analysis of items' textual contents was conducted using standard stemming tools (such as `SnowBall`[11] - based on the Porter's Algorithm [29]) and dictionaries (such as `WordNet`[12]) for the English language. In particular, we distinguish between $V_{\mathcal{W}}$, the vocabulary of (61,845) terms appearing in the WordNet dictionary, from $V_{\overline{\mathcal{W}}}$, the vocabulary of remaining (1,475,885) terms composed mostly of jargon, named entities and typos. Stop-words[13], *urls* and *e-mail* addresses were excluded and for $V_{\overline{\mathcal{W}}}$ the stemmed version of terms was kept while for $V_{\mathcal{W}}$ the terms in their base form as given by WordNet was added.

We believe that such a detailed characterization of feeds content will be beneficial to several Web 2.0 applications. For instance, knowledge regarding terms (co-)occurrence distribution could be helpful for efficient compression or indexing techniques of items textual content. The knowledge that the employed vocabulary is essentially composed of misspellings, spoken acronyms and morphological variants will greatly affect the choice of effective ranking functions for filtering the incoming items while it will enable us to devise realistic workloads for measuring the scalability and performance of Web 2.0 analysis systems (e.g., keyword tracking, buzz measuring, keyword-association in online consumer intelligence).

5.1 Term Occurrences

The global vocabulary ($V = V_{\mathcal{W}} \cup V_{\overline{\mathcal{W}}}$) extracted from the English items reaches a total number of 1,537,730 terms. Figure 6 depicts the number of occurrences of the terms belonging to $V_{\mathcal{W}}$ and to $V_{\overline{\mathcal{W}}}$ in decreasing order of their rank (frequency) in their respective vocabulary. As expected, terms from $V_{\mathcal{W}}$ are much more frequent than terms from $V_{\overline{\mathcal{W}}}$. The multiple occurrences in the items of the 5,000 most frequent terms of $V_{\mathcal{W}}$ (around 8% of its size) represent 87% of the total number of term occurrences from $V_{\mathcal{W}}$. The percentage of $V_{\mathcal{W}}$ terms

[11] snowball.tartarus.org
[12] wordnet.princeton.edu
[13] www.lextek.com/manuals/onix/stopwords2.html

appearing in the most frequent terms of V drops quickly: from 90% in the 5,000 first terms, to 78% for the next 5,000, to 50% after 20,000 terms and to only 3% for the remaining 1.5 million terms.

In the following, we are interested in characterizing formally the distribution of $V_{\mathcal{W}}$ and $V_{\overline{\mathcal{W}}}$ terms' occurrences. In this respect, Zipf's law distributions have been traditionally used in the literature [2, 17] for various text corpora:

$$f(r) = \frac{K}{r^{\theta}}$$

where r is the term rank and θ and K are constants.

However, as can be seen in Figure 6 the corresponding curve for $V_{\mathcal{W}}$ has a significant deviation from Zipf's law, *i.e.,* from a straight line in log-log scale. This deviation is smaller for the $V_{\overline{\mathcal{W}}}$ curve. Similar deviations have been already reported for web related text collections [2, 17, 1, 30, 10] and few attempts have been made to devise more adequate distributions. [18] tried to generalize the Zipf's law by proposing a Zipf - Mandelbrot distribution while [13] suggested to use a Modified Power Law distribution. Although the latter laws can approximate the slow decrease at the beginning of the curves, their tail in log-log scale is close to a straight line, which does not fit neither to our terms' occurrences nor to any of the distributions in the aforementioned studies.

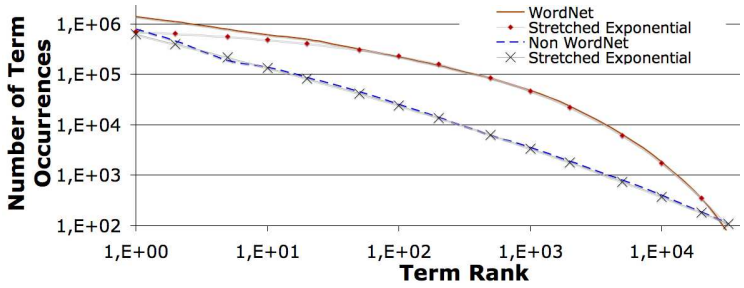

Fig. 6. Occurrences of terms from $V_{\mathcal{W}}$ and $V_{\overline{\mathcal{W}}}$

A *stretched exponential* distribution was found to better capture the tail of the distributions for both vocabularies. It is defined as follows:

$$f(r) = K \times e^{-(r/c)^{\theta}}$$

where θ expresses how fast the slope of the curve decreases, while constant c defines the curvature: for values closer to zero the distribution is closer to a straight line in log-log scale. To the best of our knowledge the stretched exponential distribution has been so far utilized to study physics phenomena but has never been associated with term distributions before, even though [12] suggests this distribution as an alternative to power laws. The constants that lead to the best fitting curves were obtained using the chi-square test (see Table 10). Lower values of chi-square indicate a better fit of the distribution to the empirical data.

	Zipf	Zipf Mandelbrot	Modified power law	Stretched Exponential
$V_{\mathcal{W}}$	2,845	2,493	966	**49.8**
$V_{\overline{\mathcal{W}}}$	11.5	11.5	9.2	**9**
Term co-ccurence	2,353	1,543	31,653	**45**

Table 10. Chi-square for 4 distributions (deg. of freedom: 210 for $V_{\mathcal{W}}$ and 200 for $V_{\overline{\mathcal{W}}}$)

Figure 7 represents the occurrences of the 32,000 most frequent term combinations in items. These combinations are all pairs of terms, not necessarily consecutive, that appear in the item's title or description regardless of the vocabulary they belong to. In contrast to [10], where the same test was performed for sponsored data, we find that this distribution does not follow the Zipf's law. In Figure 7 the curve has been fitted using the stretched exponential distribution whose chi-square value is the lowest one (see table 10).

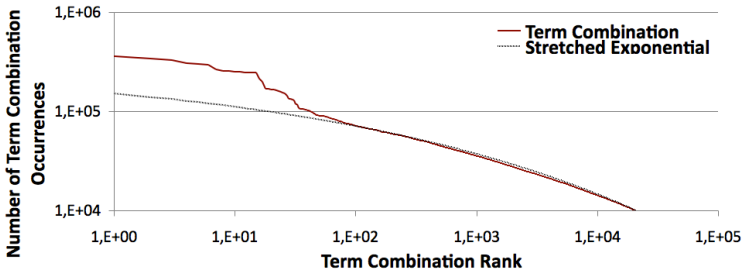

Fig. 7. Term occurrences for term combinations

The divergence of the empirical data from the stretched exponential distribution given above is limited to the 32 most frequent combinations of terms, which appear almost 4 times more than expected. These combinations are essentially due to high interests for some serials (*e.g.* like `Lost-fan`), a behavior also observed in [10].

Occurrences	1	2	3	4	≥ 5
# terms	659,159	261,906	124,450	77,814	414,401

Table 11. Terms' number per occurrences

As the number of occurrences in items varies widely from one term to another (see Figure 6) in Table 11 we provide a summary of term occurrences distribution. Terms with less than 5 occurrences represent roughly 76% of the global vocabulary size while terms with only one occurrence account for 45% of the vocabulary.

5.2 Analysis of terms

A qualitative analysis over a sample of 5,000 terms comprising the single occurrence terms that are non dictionary terms (belong to $V_{\overline{\mathcal{W}}}$) is shown in Figure 8. Since their number of occurrences is so low, one would expect that almost all of these words would be misspellings. Although, the majority (35%) of terms in this category is surprisingly named-entities, while standard mistakes represent 15% (concatenation and misspelling). There are several non English terms as well 22%, even though the feeds were declared so as to only contain English ones. Note finally, the presence of other types of terms such as wrong urls and characters, composed-words, new internet e-words and spoken acronyms.

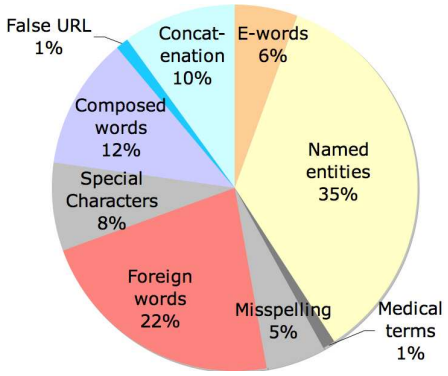

Fig. 8. Terms nature with a single occurrence

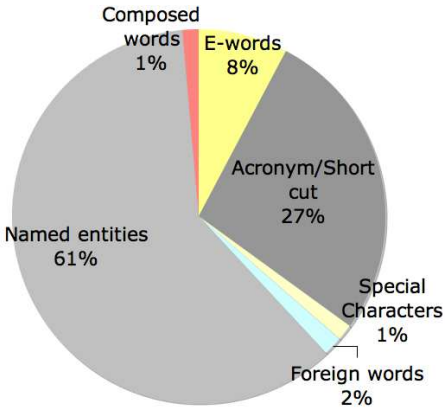

Fig. 9. Terms nature for most frequent non-wordnet terms

Figure 9 displays the frequency of each type of the 5,000 most frequent non dictionary terms (corresponding to occurences higher than 10^4 in Figure 6). Named entities are by far the most frequent terms (61%) followed by acronyms and short cuts (27%). Note also the presence of numerous new internet E-words (8%).

Clearly the language of non-authoritative web document collections on diverse topics is subject to many kinds of imperfections and errors [1], including misspellings or spoken acronyms, but also special-purpose jargon (*e.g.,* exotic technical terms in medicine or in computer science). Those imperfections are related to the informal and ephemeral use of language in the Web 2.0 era. Although, this analysis steadies a random sample for unfrequent terms and a thorough sample for most frequent terms, we can say that named-entities have a real impact on non vocabulary terms extraction, however we need to pay attention to E-words and shortcuts which seem to take a non-negligible place.

5.3 Vocabulary size

Figure 10 illustrates the size growth wrt the number of items progressively stored in our warehouse for the two vocabularies of terms. Clearly, the size of $V_{\overline{\mathcal{W}}}$ evolves much faster than that of $V_{\mathcal{W}}$: at the end of the 8-month period, it is 24 times bigger than the size of the former ($|V_{\mathcal{W}}| = 61{,}845$ while $|V_{\overline{\mathcal{W}}}| = 1{,}475{,}885$).

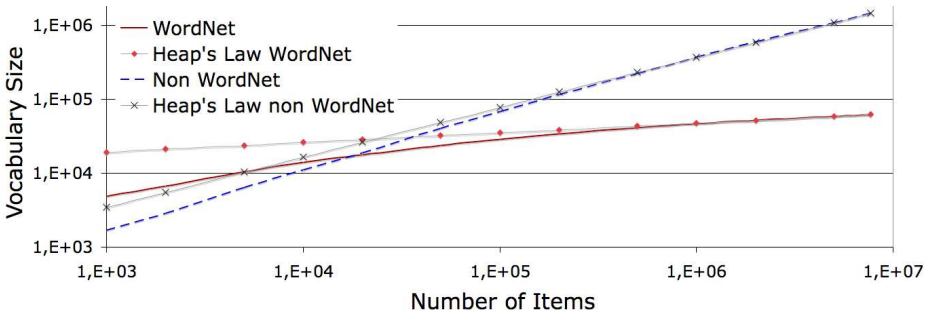

Fig. 10. Evolution of the two vocabularies

For $V_{\mathcal{W}}$, a rapid increase is observed for the first 140,000 items to reach the 31,000 most frequent terms of $V_{\mathcal{W}}$, while the size barely doubles up to the end, with the inclusion of less frequent terms. On the other hand, $V_{\overline{\mathcal{W}}}$ has a smoother growth: for 140,000 items $V_{\overline{\mathcal{W}}}$ comprises 88,000 terms, *i.e.* only 6% of the total $V_{\overline{\mathcal{W}}}$ size which reaches almost 1.5 million terms after processing all items.

The vocabulary size growth is usually characterized by a Heap's law distribution [2, 17] (see Figure 10), defined as follows:

$$|V(n)| = K \times n^{\beta}$$

where n is the number of collected items while K and β (taking values in $[0, 1]$) are constants depending strongly on the characteristics of the analyzed text corpora and on the model used to extract the terms [30]. β determines how fast the vocabulary evolves over time with typical values ranging between 0.4 and 0.6 for medium-size and homogeneous text collections [2]. [30] reports a Heap's law exponent lying outside this range ($\beta = 0.16$) for a 500 MB collection of documents from the Wall Street Journal. Table 12 specifies the constants chosen for Heap's laws approximating the global vocabulary growth as well as $V_{\mathcal{W}}$ and $V_{\overline{\mathcal{W}}}$.

Clearly, the Heap's law exponent (β) of the global vocabulary is affected by the evolution of $V_{\overline{\mathcal{W}}}$ rather than by $V_{\mathcal{W}}$ whose size is significantly smaller (attributed to the slow acquisition of less commonly used terms). The exponent for $V_{\overline{\mathcal{W}}}$ (0.675) slightly higher than those reported in the literature [2, 17] indicates a faster increase of the vocabulary size due the aforementioned language imperfections of the items in our testbed. This behavior is related to texts that are sent

	$\lvert V_{\mathcal{W}} + V_{\overline{\mathcal{W}}}\rvert$	$\lvert V_{\mathcal{W}}\rvert$	$\lvert V_{\overline{\mathcal{W}}}\rvert$
K	51	7,921	33
β	0.65	0.13	0.675

Table 12. Heap Laws constants

once by publishers and never corrected (i.e. in a case of misspelling), which leads to a higher number of errors (i.e. for $V_{\overline{\mathcal{W}}}$) and therefore a faster than expected vocabulary evolution. It should be stressed that this behavior is also exhibited by the evolution of vocabularies related to other user-generated textual content (such as web queries [31, 23]). Finally, the high Heap's law coefficient (K) for $V_{\mathcal{W}}$ is explained by the rapid vocabulary growth in the beginning due to the fast acquisition of the very popular terms.

5.4 Rank variation

Since the number of occurrences of a term evolves over time as new items arrive, their corresponding ranking could be also subject to change. Figure 11 illustrates the average variation of ranks in the global vocabulary during the last month of our study. A rank variation is measured by the difference of a term rank between week t-1 and week t. Obviously, this variation is proportional to the rank, $D = \alpha \times r$, *i.e.* the less frequent terms are more likely to change their rank compared to the vocabulary of the previous week. The observed gradient α of the distance D after 8 month is equal to 0.035 for $V_{\overline{\mathcal{W}}}$ and 0.013 for $V_{\mathcal{W}}$.

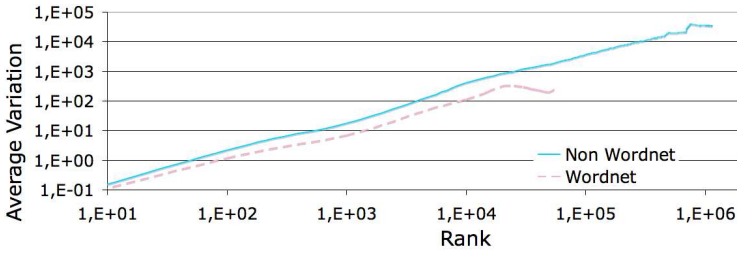

Fig. 11. Cumulative rank distance in V (last month)

We are finally interested in studying the weekly rank variation for specific rank ranges of $V_{\overline{\mathcal{W}}}$ terms for the 8 months of our study using the *Spearman* metric [11]. The *Spearman* metric is the sum of rank variations for a range of ranks from a week to another:

$$S(t) = \sum_{i=1}^{n} |r(t)_i - r(t-1)_i|$$

where t is a given week, $r(t)_i$ and $r(t-1)_i$ are the rank of the term i at week t and $t-1$ respectively. Figure 12 depicts the weekly *Spearman* value for three classes of terms' ranks: (a) highly-frequent terms (ranks between 1 and 250), (b) commonly-used terms (10,000 to 10,250); and (c) rare ones (500,000 to 500,250). Note that the reference vocabulary for this experiment is chosen as the vocabulary obtained after the first three months of items' acquisition.

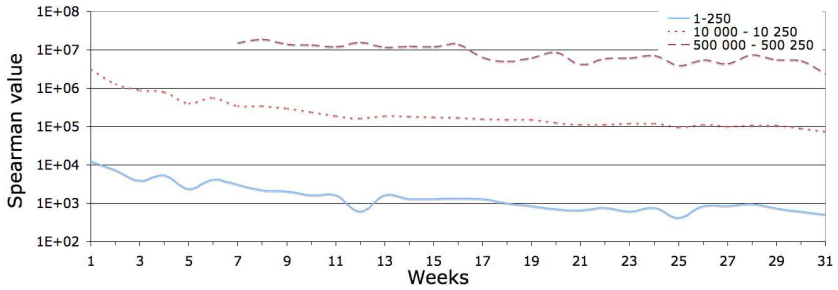

Fig. 12. Weekly Spearman Variation for three rank ranges of $V_{\overline{W}}$ terms

Surprisingly enough, we have observed that the ranking of vocabulary terms does not significantly vary during the 8-month period of the study for highly-frequent terms. As a matter of fact, the sum of rank variations of highly-frequent terms is 4 orders of magnitude less than the corresponding sum for rare terms (where commonly-used terms lie in between). We notice that the Spearman value stabilizes after a 20-week period for all the rank ranges. This can be justified by the fact that the number of occurrences of each term after this period decreases the impact of incoming terms on the rank.

So the vocabulary analysis reveals that RSS vocabulary includes a large amount of imperfections and errors (*e.g.* E-words, misspells, concatenations) which may have an impact on the design and performance of RSS filtering systems on textual content. We observe that the growth of the global vocabulary is bounded by a Heap law leading to a cleaning process of the vocabulary based on terms occurrences. On the other hand, some text-based systems (*e.g.* indexes) are based on rank. Our analysis stresses the strong rank evolution of most uncommon terms which rank changes up to 500 places from a week to an other. This changes may degrades the performances of structures based on ranks.

6 Conclusion

In this paper we presented an in-depth analysis of a large testbed of 8,155 active RSS feeds comprising 10.7 million million of items collected over a 8 month period campaign. We proposed a characterization of feeds according to their publication rate and temporal variability in which we identified three different

activity classes. In addition, we focused on RSS items length, structure and replication rate across feeds. Last but not least, a detailed study of the vocabulary employed in a significant subset of the RSS items written in English was conducted along three lines: occurrence distribution of the terms, evolution of the vocabulary size and evolution of term ranks.

We believe that such a precise and updated characterization of feeds behavior and content is beneficial to different Web 2.0 applications. For example, knowing that highly active feeds are more predictable in their publication behavior, would guide resource allocation mechanisms to be tied to the feeds category. Furthermore, being aware that the most popular RSS/Atom elements are textual with an average length smaller than blogs and web pages would impact the design of content-based RSS subscription languages as well as of the subsequent Pub/Sub indexes. Finally, taking into account that the employed vocabulary is composed of misspellings, spoken acronyms and morphological variants will greatly affect the choice of effective ranking functions for matching approximately user subscriptions to incoming items while concrete measures regarding vocabulary size and terms occurrence distribution allow to devise realistic workloads for measuring the scalability and performance of Web 2.0 retrieval and analysis systems (e.g., for keyword tracking, buzz measuring, or keyword-association in online consumer intelligence applications).

References

1. Ahmad, F., and Kondrak, G. Learning a Spelling Error Model from Search Query Logs. In *EMNLP* (2005).
2. Baeza-Yates, R. A., and Ribeiro-Neto, B. A. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
3. Bouras, C., Poulopoulos, V., and Tsogkas, V. Creating Dynamic, Personalized RSS Summaries. In *ICDM* (2008), pp. 1–15.
4. Choi, S.-W. Some statistical properties and zipf's law in korean text corpus. *JQL 7*, 1 (2000), 19–30.
5. Haghani, P., Michel, S., and Aberer, K. The gist of everything new: personalized top-k processing over web 2.0 streams. In *CIKM* (2010), ACM, pp. 489–498.
6. Hatzigeorgiu, N., Mikros, G., and Carayannis, G. Word length, word frequencies and zipf's law in the greek language. *JQL 8*, 3 (2001), 175–185.
7. Hristidis, V., Valdivia, O., Vlachos, M., and Yu, P. S. A System for Keyword Search on Textual Streams. In *SDM* (2007).
8. Hu, C.-L., and Chou, C.-K. RSS Watchdog: an Instant Event Monitor on Real Online News Streams. In *CIKM* (2009), pp. 2097–2098.
9. Irmak, U., Mihaylov, S., Suel, T., Ganguly, S., and Izmailov, R. Efficient Query Subscription Processing for Prospective Search Engines. In *USENIX* (2006), pp. 375–380.
10. König, A. C., Church, K. W., and Markov, M. A Data Structure for Sponsored Search. In *ICDE* (2009), pp. 90–101.
11. Kumar, R., and Vassilvitskii, S. Generalized distances between rankings. In *WWW* (2010), pp. 571–580.

12. Laherrère, J., and Sornette, D. Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *Eur. Phys. J. B 2*, 4 (1998), 525–539.
13. Lambiotte, R., Ausloos, M., and Thelwall, M. Word Statistics in Blogs and RSS Feeds: Towards Empirical Universal Evidence. *CoRR* (2007).
14. Levering, R., and Cutler, M. The portrait of a common html web page. In *ACM Symp. on Document Engineering* (2006), pp. 198–204.
15. Liu, H., Ramasubramanian, V., and Sirer, E. G. Client Behavior and Feed Characteristics of RSS, a Publish-Subscribe System for Web Micronews. In *IMC* (2005), pp. 3–3.
16. Ma, S., and Zhang, Q. A Study on Content and Management Style of Corporate Blogs. In *HCI (15)* (2007), pp. 116–123.
17. Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
18. Montemurro, M. A. Beyond the zipf-mandelbrot law in quantitative linguistics. *Physica A 300*, 3-4 (2001), 567–578.
19. Petrovic, M., Liu, H., and Jacobsen, H.-A. CMS-ToPSS: Efficient Dissemination of RSS Documents. In *VLDB* (2005), pp. 1279–1282.
20. Pitoura, T., and Triantafillou, P. Self-join size estimation in large-scale distributed data systems. In *ICDE* (2008), pp. 764–773.
21. Press, O. U. Rt this: Oup dictionary team monitors twitterer's tweets, June 2009.
22. Roitman, H., Carmel, D., and Yom-Tov, E. Maintaining dynamic channel profiles on the web. *VLDB 1*, 1 (2008), 151–162.
23. Schmidt-Maenz, N., and Koch, M. Patterns in search queries. In *Data Analysis and Decision Support* (2005).
24. Sia, K. C., Cho, J., and Cho, H.-K. Efficient monitoring algorithm for fast news alerts. *TKDE 19* (2007), 950–961.
25. Silberstein, A., Terrace, J., Cooper, B. F., and Ramakrishnan, R. Feeding frenzy: selectively materializing users' event feeds. In *SIGMOD* (2010), pp. 831–842.
26. Taddesse, F. G., Tekli, J., Chbeir, R., Viviani, M., and Yetongnon, K. Semantic-Based Merging of RSS Items. *WWW 13*, 1-2 (2010), 169–207.
27. Thelwall, M., Prabowo, R., and Fairclough, R. Are Raw RSS Feeds Suitable for Broad Issue Scanning? A Science Concern Case Study. *JASIST 57*, 12 (2006), 1644–1654.
28. Van Kleek, M., Moore, B., Karger, D. R., André, P., and schraefel, m. Atomate It! End-User Context-Sensitive Automation Using Heterogeneous Information Sources on the Web. In *WWW* (2010), pp. 951–960.
29. van Rijsbergen, C., Robertson, S., and Porter, M. New Models in Probabilistic Information Retrieval. *London: British Library Research and Development Report 5587* (1980).
30. Williams, H. E., and Zobel, J. Searchable words on the Web. *JODL 5*, 2 (2005), 99–105.
31. Zien, J. Y., Meyer, J., Tomlin, J. A., and Liu, J. Web Query Characteristics and their Implications on Search Engines. In *WWW* (2001).