On Estimating The Geographic Distribution Of Social Media

Matthew Hurst
Nielsen BuzzMetrics
matthew.hurst@buzzmetrics.com

Matthew Siegler
Nielsen BuzzMetrics
matthew.siegler@buzzmetrics.com

Natalie Glance
Nielsen BuzzMetrics
natalie.glance@buzzmetrics.com

Abstract

Many social media platforms allow the user to provide profile information. This information is generally presented in a semi-structured manner either on a profile page or on the weblog home page itself. This paper describes a novel wrapper induction method that extracts profile data. Our ultimate goal is to estimate the geographic distribution of weblog authors and to that end we provide an analysis of the location information discovered for each author in a large database of weblog posts.

Keywords

weblogs, geolocation

1. Introduction

The blogosphere is global. But what is the distribution of authors like? How many come from the United Kingdom? the US? France? In order to answer that, we need to estimate the size of the blogosphere and the distribution of the location of authors. As an initial step towards that eventual goal, we present in this paper a study on the distribution of the location of bloggers for a set of hosts.

2. Profile Data in Social Media

Many social media platforms, both hosted and self-hosted, allow for and even require the creation of profiles. The fields in a profile may include, for example, the gender of the author, the age of the author, the location of the author, their hobbies or interests and so on. Some of this data may be free text whereas others may be structured in ways specific to the platform.

Some or all of this data may be made visible by either the platform or the author. In many cases, the data will be available on a specific profile page which may be linked to from appropriate areas of the platform. In other cases, the data may be published only on the blog or board site with no specific profile page. Finally, there are cases where a profile page is available, but a summary of the data is also made available in other locations.

3. Methodology

We outline our approach here in terms of weblogs and weblog authors. Later, we will analogue results for message boards.

In the course of collecting blog post data we have also acquired a parallel collection of blog profile page information.

From our collection of authors, we use a number of readily available data.

- The number of authors.
- The distribution of authors over blogging platforms.
- The number of authors associated with each platform.
- The distribution of authors over countries according to information published in profile pages and elsewhere on the blog.
- The distribution of authors over countries per platform.

From these values, we can start to build a picture of the distribution of bloggers *for posts in our blog post collection*. Note that we view this work as part of a larger project to establish various parameters describing the true population of the blogosphere.

4. Issues and Assumptions

There are some clear problems with the approach that we outline in this paper. The two most important are

- Self-selection: authors from different countries, different linguistic communities and different social networks or platforms may elect to disclose information in different ways. For example, if Russians were more likely to disclose their location than Chinese our results would be skewed towards the Russian population.
- The size of the blogosphere. We are interested in making projections based on available data. However, as we don't know the size of the entire blogosphere, it would be imprudent of us to predict the number of bloggers from any different country. For the purpose of this work, which is preliminary, we avoid any absolute predictions by stating results in percentages. However, it can't be ignored that the pool from which we draw bloggers may be skewed by the *absence* of data from certain platforms.

5. Profile Understanding

Wrapper induction, in its most general sense, is the creation of a model which, when parameterized by an object within an appropriate system yields a structured view of the object. The most common form of wrapper application takes

ICWSM'2007 Boulder, Colorado, USA

a document, specifically web documents written in HTML, and provides access to parts of the document in a structured manner - essentially providing a type to these pieces of content. For example, a wrapper might give access to the person name field in a job seekers posting; or a headline field in a news article.

The first example provides a semantic wrapper - it essentially says 'this part of the document contains text which refers to an address.' The second example is an example of structure wrapping - it provides access to the logical structure of the document.

We can characterise wrapper generation systems along a number of dimensions (c.f. [2] for a similar characterisation).

- Static versus Dynamic. A static wrapper is stored and retrieved for application. A dynamic wrapper is created at interpretation time.
- Inference versus Description. An inferred wrapper uses some form of training information to learn the wrapper. A manual system is simply an interface to allow a person to describe the wrapper.

Within the space of inference systems, further specialization exists in the manner in which inference is carried out, the type of training data and so on. In the context of this paper, the most important distinction lies between human generated training examples (in which a person annotates an example of the target page type with desired extractions) and model driven analysis (in which some automated process is provided which is capable of annotating the page with certain types of candidates).

¿From the literature, [1] provides an example of a static inference system, [3] provides an example of a dynamic inference system. The work presented in this paper is probably best compared to that in [2] and [3]. The former aims to understand the meaning in repeated structure and to model that structure. The later is more constrained in scope and uses a model of the structure of weblogs to create a wrapper for any weblog.

The wrapper system described in this paper is a dynamic, model driven inference system

The structure of documents found on Social Media platforms is generally very simple. A weblog, for example, consists of a series of posts each with a handful of fields (title, body, footer containing date, author name, comments, etc.). A message board has a similarly restricted page structure. At the site level, again, there is plenty of repeated structure which can be adequately described by a relatively small abstract model.

This type of data, therefore, is certainly attractive to model driven learning.

5.1 Approach

It is instructive to look at a few simple examples of profile data presentation. LiveJournal and Xanga both use a simple table based layout. The difference being that LiveJournal (shown below) uses some additional face information (`b`).
LiveJournal

```
<tr>
  <td><b>Name:</b></td><td>Matthew Hurst</td>
</tr>
<tr>
  <td><b>Country:</b></td><td>United States</td>
</tr>
```

MySpace, on the other hand, uses a single table cell, key values are in bold (`b`) and entries are separated by a line break (`br`).

```
<tr>
  <td>
    <b>Gender</b>: Male<br>
    <b>Status</b>: Married<br>
    <b>Age</b>: 27<br>
  </td>
</tr>
```

In these two examples, the rendering is similar but the encoding is quite different. In the first case the key and value are structurally related by being sequenced cells in the same table row. In the second example, the key and value are related by being rendered on the same line (or, to strictly interpret the HTML, by not being forced to be on different lines). In addition - to making observations about the relationships between the two nodes in a key/value pair - we also get a strong hint regarding the type of data present by the fact that there are repeated patterns in the data. In the first it is the presence of pairs of table cells, in the second it is the presence of pairs of text nodes, one of which (the first) is emboldened.

As a final example, consider the following from a message board (ezboard).

```
<div id="personalinfo">
  <h2>My Personal Information</h2>
    <div id="firstname">
      <span class="title">First Name ::</span>
      <span class="value">Matthew</span>
    </div>
    <div id="lastname">
      <span class="title">Last Name ::</span>
      <span class="value">Hurst</span>
    </div>
    <div id="age">
      <span class="title">Age ::</span>
      <span class="value">27</span>
    </div>
  </div>
</div>
```

Here, there is extensive use of div and span encoding - a different approach to the previous examples.

The insights that motivate the design of the wrapper induction system described here, then, are as follows:

Content The data presented in profiles generally includes elements from a small pool of core values such as gender, location and age.

Alignment Key/value pairs are aligned horizontally.

Global Structure Multiple key/value pairs have the same global structure (they keys are positioned similarly, the values are positioned similarly).

Local Structure Multiple key/value pairs have the same local structure (the relationship between the key and the value is the same for all pairs).

Given these assumptions, the algorithm introduced in this paper works in the following manner.

1. The text nodes in a document are extracted.
2. Using an examplar set of key terms (e.g. `name`, `age`, `gender`), a seed set of text nodes is extracted.
3. The set of all *similar* text nodes (see later for definition of *similar*) matching any example in the seed set is extracted. This set forms the working set.
4. The working set is then partitioned into sets of nodes which all match each other.

The partitions are then processed to discover repeated structures of key/value pairs.

1. The scope of the partition is computed. The scope is the longest path which contains all the text nodes.
2. A structural relation is computed for the partition (see later).
3. All pairs of text nodes for which the structural relationship holds in the partition are extracted as key/value pairs.

A complete formal description of the algorithm is beyond the scope of this paper. Therefore, we present below detailed descriptions of the core data structures and component algorithmic elements.

5.2 Definitions

T is the set of text nodes in the DOM. If x is the path of DOM nodes from the root of the document to a text node, then d, the **text node descriptor**, is a tuple $< p, F, n, t >$ where:

p is the path - a sequence of nodes that represent the path to this node excluding the node in F.

F is the font description - a set of nodes which represents the accumulated face and font describing html elements.

n is the node - the text node being modeled.

t is the text model - a representation of the text which has the following structure (prefix, suffix, content).

Two text nodes are determined to be *similar* if their paths are identical, their font descriptions are identical and their text models are identical. Text models are identical if the prefix and suffix are identical. Thus `table/tr/td/b/'Name'` is *similar* to `table/tr/b/td/'Gender'`.

A partitioning of D is defined as follows: Each subset must contain members which are all similar to each other.

The scope of a set of descriptors (scope(D)) is a descriptor representing the longest common subsequence of nodes starting at the document root.

An *SREL* (structural relationship) is a relationship between two nodes in a document. For example, there is an obvious structural relationship between text in the same row of a table or the same column. As a proxy for this, we can look at the DOM encoding of the document layout and infer structural relationships in the document via relationships in the DOM. To continue the example, two text nodes below the same TR node are structurally related.

The method *computeSREL*(.) takes two text node descriptors and tests a number of conditions in sequence. The tests are ordered by hand so that the first match that returns - that is to say, the first detected structural relationship - is the 'best' one.

The matching algorthim is outlined below.

```
computeSREL(D1, D2)
  l → |D1| − 1        //l is the length of the D1's path
  for(p=l; p> =0; p–)
    node ← D1[p]    //node becomes the p^th node in
                    //D1's path
    if(D2 doesn't contain node)
      continue
    //otherwise, we start
    //to perform tests
    if(node is a p node)...
    if(node is a tr node)...
    if(node is a td node)...
    ...
```

Each test, which is conditioned on the type of the node that the two paths share. For example, if the two text nodes are in the same paragraph node (`p`), then they have the SREL_P relation only if there is no explicit line break between them and that they are ordered left to right. This captures, cases like the following.

```
<p>
  <b>Location</b>: Elie, Fife
</p>
```

Where *computeSREL* would return SREL_P if given the text node descriptors `Location` and `: Elie, Fife` in that order.

5.3 Profile Interpretation

Once we have extracted the fields from a profile, we must interpret them. For location information, this means mapping from strings to a set of symbols representing countries. There are a number of standards that are in common use for representing geographic information. For country level data, we use ISO 3166.

The strings extracted from profile pages are quite varied and include the following elements: country name, state or province, city name. Our system currently takes a pragmatic approach to interpretation. If there is an identifiable country name, then that is used. If there is a US state, then we infer US as the country. If there is a populous city (e.g. a capital) then the nation for that city is taken. Our database of names includes many synonyms and international encodings, though we do not claim that it is exhaustive.

We acknowledge that even after we have extracted a country correctly, our ultimate result - the location of the author - may be incorrect. There are two main reasons for error. Firstly, the author may simply be providing misinformation. Secondly, the precise semantics of the location field may not be entirely clear. Does it capture the home town of the author or the present location? For example, consider a blogger from London living as a student in the US. If they start blogging in the US, which location do they enter?

Result	Count
OK	144
Extraction Error	42
Foreign	18
Interpretation Error	3
Total	207

Table 1: *Result classes for location extraction task.*

Host	Count	Percent
blog.myspace.com	26	12.6%
www.xanga.com	7	3.4%
blogspot.com	5	2.4%
www.blogger.com	2	1.0%
spaces.msn.com	1	0.5%
spaces.live.com	1	0.5%
Total	42	

Table 2: *Distribution of extraction errors.*

Currently, we don't attempt to resolve these issues.

5.4 Evaluation

We selected 207 blogs distributed over hosts as shown in Table 3. We manually labeled the results of the location extraction system with the following classes:

OK the system correctly identified the country declared by the author.

Extraction Error the system failed to extract anything.

Foreign the system failed to extract anything and the page was in a foreign language (which has the potential to impact the recall as the key names (location, etc.) are drawn from a non-English set.

Interpretation Error the system successfully extracted the text but a mapping wasn't found to a valid country.

The distribution of results is shown in Table 1. Of the extraction errors, the distribution over domains is shown in Table 2. The distribution over domains in the original data set is shown in Table 3.

The extraction error results indicate that the main problem is caused by myspace blogs. An inspection of the system revealed that the addition of a span based structural relationship solved the problem. It is important to note that additional structural relations have been added conservatively. Rather than create a relation for any node as long as the two descriptors share it, we opted to extend a closed set of relations on a case by case basis to ensure that the system didn't over generalize.

Host	Count
xanga.com	43
spaces.msn.com	37
blogger.com	36
blog.myspace.com	29
spaces.live.com	25
blogspot.com	23
livejournal.com	10

Table 3: *Test data set distribution.*

The next main class of error was that of errors foreign data. Here, there are two issues:

- the presentation of profile information with key/value pairs in a non English language (requiring the addition of new key candidate values).
- the interpretation of location information. This is handled currently, to a large extent, via the addition of new terms in the country name synonym list. However, we must also anticipate highly localized content. We already do this for US based systems, where such locations as 'Florida' or 'Miami' can be interpreted trivially assuming the US is in context. However, we may well find locally scoped blog hosts from other parts of the world where we have to rely on similar assumptions.

6. Analysis

The basic idea is as follows. We take some sample of blogs and, through either their profile page or through analysing their blog home page we discover their self declared location. We now have two sets of blog author identities. The larger set is the original sample. The smaller subset is that set for which we successfully extracted and interpreted a location.

We assume that the distribution of authors to countries found in blogs is different between different hosts. For example, the ratio of US, Russian and Chinese bloggers on LiveJournal compared with the ratio on Spaces is quite different (as supported by [4]). Consequently, we project the number of authors by summing the projections per domain. For the set of countries C, and the set of domains D, for a given country n, c(l, d) is the count of authors for country l in domain d, $a(d)$ is the total count of authors in domain d.

$$\sum_{d \in D} \left(\frac{c(n,d)}{\sum_{l \in C} c(l,d)} \cdot a(d) \right) \tag{1}$$

Compare this with the independent project (in which it assumed that the distribution of authors is not dependent on the domain).

$$\frac{\sum_{d \in D} c(n,d)}{\sum_{d \in D} c(d)} \cdot \sum_{d \in D} a(d) \tag{2}$$

Figure 1 shows results for our domain-dependent approach. These results can be compared with those from the domain-independent method in Figure 2. Here, the closer to the diagonal a country is, the less dependent the data is on a domain.

We carried out a similar analysis of authors in our message board content system. The blog system from which we drew author data has a collection mechanism that relies on the ping infrastructure of the blogosphere. Our message board system, by contrast, has been populated in a different manner due to the lack of such an infrastructure. Consequently, the data in our message board content system is from almost exclusively English language boards. This provides a great opportunity to look at the geographic distribution of participants in the English language boardscape. Figure 3 shows the distribution of locations for authors on message boards based

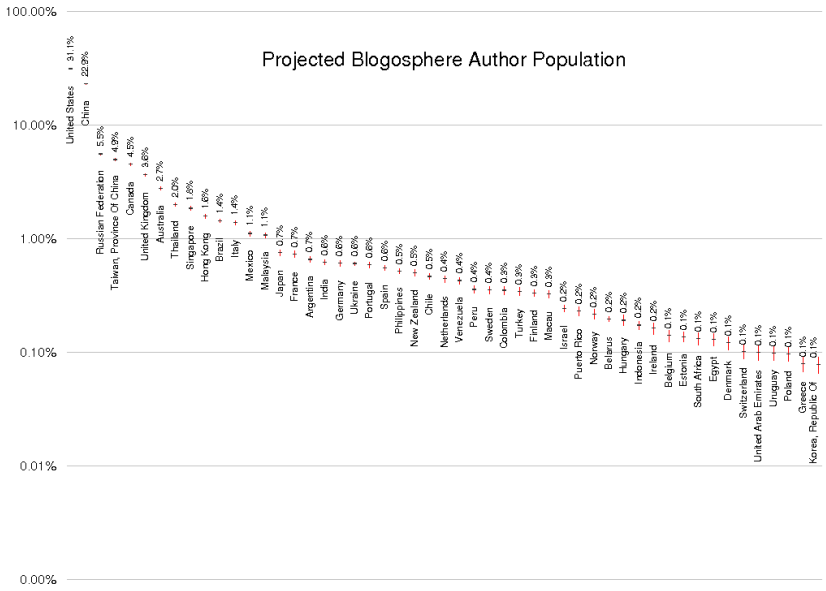

Fig. 1: *Distribution of weblog authors projected with dependence on hosting domain.*

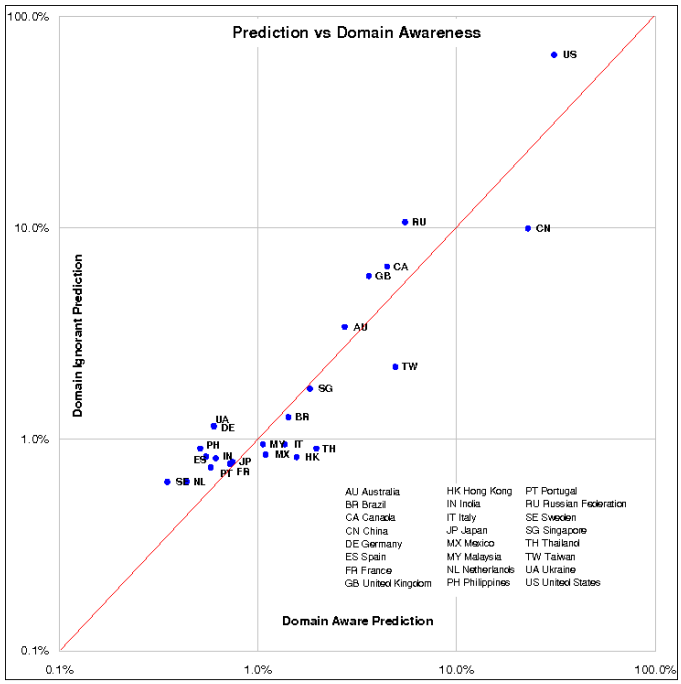

Fig. 2: *Comparison of dependent and independent projections for weblog authors.*

on our prediction methodology. This is then compared with the independent approach in Figure 4.

There are some key differences between the two projections. Perhaps the most immediate to address concerns the error bars. The error bars (which indicate the 95 % confidence ranges) for the blog projection are far smaller than those for the message boards. This is due in part to there being, for blogs, a small number of large domains with many populated profiles. This is in contrast to the boardscape where we see many more domains with smaller numbers of authors and fewer examples of geolocation in the profiles.

7. Comparison

We can compare the results from our projections with statistics from Technorati's language breakdown for posts and with the language breakdown by blog for our own blog data, as shown in Table 4.

Sifry's State of the Blogosphere ([6]) provides statistics of language *per post* from the Technorati weblog index. Comparing the Technorati language breakdown with our author data is not straightforward. Firstly, Technorati's data is over posts, not authors, and, secondly, Technorati's index contains a noticable amount of non-post data (including weblog home pages and some non-weblog content). However, along with expected similarities (the dominance of English language and English speaking locations), there are some notable differences. In particular, our projections suggest that Chinese and Russian should appear prominently in the language based segmentation.

The match between geolocation and language improves when we compare location breakdown with the language breakdown for blogs collected by BlogPulse in October 2006. This is not surprising, as the BlogPulse blog data was used as a source set of blog urls for harvesting blog author profiles. Thus, we find English, Chinese and Russian languages to be strongly represented as the location segmentation implies. However, the relative percentages for languages vs. location are significantly different. For example, 42 % of bloggers claim a location in a predominantly English-speaking country, but 45 % of blogs are written in English. Likewise, 38 % of bloggers claim a Chinese location (mainland China or Taiwan), but only 20 % are written in Chinese, while 5.5 % of bloggers claim to live in Russia, but only 4.3 % are written in Russian. This seems to indicate that some bloggers living outside the U.S. choose to write in English, either because English is viewed as a globally understood language or because these bloggers are expatriates from an English-speaking country. We plan to investigate further the breakdown by language for given locations in future work.

Perhaps the most dramatic, or suspicious, difference found here is for Japanese blogs. Both the Technorati and BlogPulse language numbers indicate a far larger Japanese blogosphere than our predicted numbers (33 % and 12.7 % versus 0.7 % by geolocation). The discrepancy occured because our system fails to interpret a majority of Japanese locations due to two main reasons. First, because many Japanese bloggers use Japanese-based blogging systems, seed terms like 'age' and 'location' are likely to be in Japanese, not English, which our system does not recognize. Second, the profile extraction system may fail to interpret locations in Japan that don't include country information nor match populous areas. (Note that these problems exist to a greater or lesser extent for other country locations, depending on the prevalence of language-specific blog hosts and their popularity and on the ambiguity in mapping locations to countries.)

Language	% Posts (Technorati)	% Blogs (BlogPulse)
English	39 %	45 %
Japanese	33 %	12.7 %
Chinese	10 %	20.2 %
Spanish	3 %	2.7 %
Italian	2 %	3.3 %
Russian	2 %	4.3 %
Portuguese	2 %	0.9 %
French	2 %	1.7 %
German	1 %	0.6 %
Farsi	1 %	1.4 %

Table 4: *Distribution of languages for posts in Technorati versus for blogs in BlogPulse (for Octboer 2006).*

It should also be underlined (as Sifry explained in [5] in the context of Technorati's blog index) that the BlogPulse blog data grossly under-represents the Korean blogosphere and the French blogosphere, because large Korean blog providers (like Cyworld or Planet Weblog) and the predominant French blogging system, Skyblog.com, are not being indexed by BlogPulse currently. The reason for their absence is because they are walled gardens. It is also likely that BlogPulse does not have the same coverage for the Japanese blogosphere, a country for which Technorati has made special efforts.

Hurst's study 24 hours of blog pings was a first foray into collecting and analyzing the locations of bloggers [4]. This work showed that the distributions of bloggers that provided location information on different platforms (specifically Microsoft's Spaces and Google's Blogspot) were very different. It was also revealed that the number of bloggers that disclosed their location information was platform-dependent, with Spaces bloggers disclosing location 81.65 % of the time compared with 34.26 % on Blogspot. The larger Japanese blogosphere is also supported by the numbers published in [4].

8. Conclusion

This paper reports ongoing work with two main goals: providing a database of authorial information for social media authors; understanding the distribution of the location of authors. We have presented a novel algorithm for wrapper induction from weblog and message board profile pages and described and discussed current results drawn from our database of author locations. Significantly, we have shown results that demonstrate a clear geographic bias in both weblog hosting systems and message board systems.

References

[1] W. W. Cohen, M. Hurst, and L. S. Jensen. A flexible learning system for wrapping tables and lists in html documents. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 232–241, New York, NY, USA, 2002. ACM Press.

[2] B. Gazen and S. Minton. Autofeed: an unsupervised learning system for generating webfeeds. In *K-CAP '05: Proceedings of the 3rd international conference on Knowledge capture*, pages 3–10, New York, NY, USA, 2005. ACM Press.

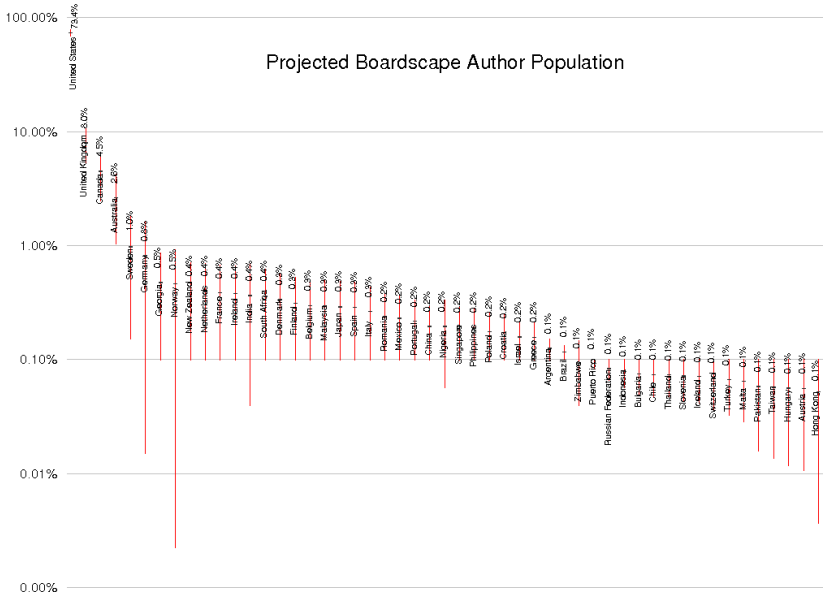

Fig. 3: *Distribution of message board authors projected with dependence on hosting domain.*

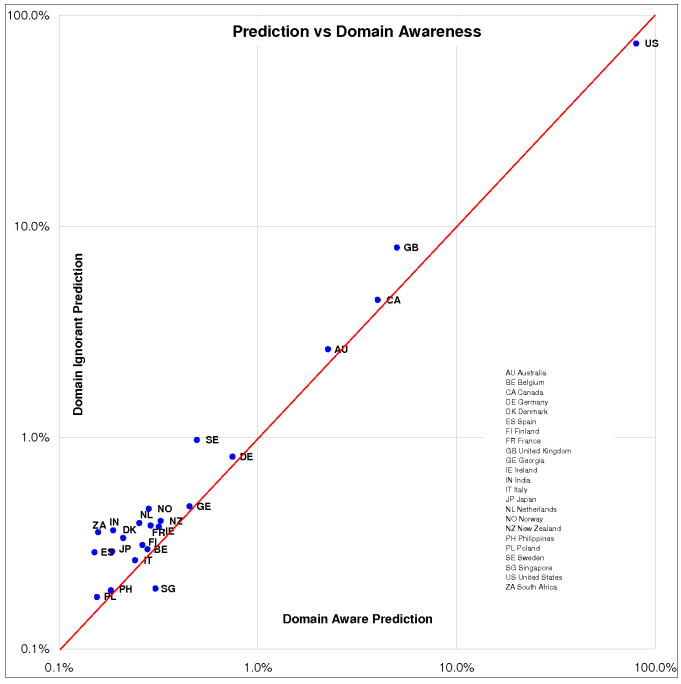

Fig. 4: *Comparison of dependent and independent projections for message board authors.*

[3] N. Glance. Indexing weblogs one post at a time. In *AAAI Spring Symposium: Computational Approaches to Weblog Analysis*, 2006.

[4] M. Hurst. 24 hours in the blogosphere. In *AAAI Spring Symposium: Computational Approaches to Weblog Analysis*, 2006.

[5] D. Sifry. State of the blogosphere, april 2006 part 2: On language and tagging.

[6] D. Sifry. State of the blogosphere, october 2006.