A Query-Driven Characterization of Linked Data

Harry Halpin
Institute for Communicating and Collaborative Systems
University of Edinburgh
2 Buccleuch Place
Edinburgh, United Kingdom
H.Halpin@ed.ac.uk

ABSTRACT

Due to the Linked Data initiative, the once unpopulated Semantic Web is now rapidly being populated with millions of facts stored in RDF. Could any of this data possibly be interesting to ordinary users? In this study, we run queries extracted from a query log from a major hypertext search engine against a Semantic Web search engine to determine if the Semantic Web has anything of interest to the average Web user. There is indeed much Semantic Web information that could be relevant for many queries for entities (like people and places) and abstract concepts, although these possibly relevant results are overwhelmingly clustered around DBPedia. We present an empirical analysis of the results, focusing on their major sources, the structure of the triples, the use of various RDF and OWL constructs, and the power-law distributions produced by both the URIs that serve Linked Data and the URIs in the triples themselves. The issue of 303 redirection and URI identity is given in-depth treatment.

Categories and Subject Descriptors

H.3.d [**Information Technology and Systems**]: Metadata

General Terms

Experimentation

Keywords

Linked Data statistics, query logs, information retrieval,power law

1. INTRODUCTION

What are the characteristics of the Linked Data in the wild? There are two primary questions we are hoping to answer. First, has Linked Data changed from earlier 'first generation' Semantic Web efforts? Second, is there anything worth finding for ordinary users in Linked Data? Only a moderately large-scape sampling and analysis of Linked Data can answer this central question. Our method of investigation is to inspect what information needs *actual* users are expressing via using a hypertext search engine, and then use a sample of these queries to determine if Linked Data can satisfy these information needs. We present an analysis of a search-engine query log from a major hypertext search engine, Microsoft's *Live.com*, and use this query log to sample Linked Data. As an added benefit, such an empirical analysis can prove or disprove some widely held assumptions, such as whether or not there is an endemic over-use of `owl:sameAs` and whether or the Linked Data best practice recommendation of `303` redirection is being followed.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.
LDOW 2009, April 20, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. PREVIOUS WORK

For the first-generation of the Semantic Web, there was very little data-driven analysis of the ontologies, primarily because so few were actually in existence. The first large-scale analysis of the Semantic Web was done via an inspection of the index of Swoogle by Ding and Finin [16]. Ding and Finin first estimated the size of the Semantic Web to be in 2006 4.91 million Semantic Web documents via searching Google for the media type `application/rdf+xml` [16]. As this might not include data that is hosted using the wrong media type, they estimated, using Google to include all FOAF files served as HTML and RSS 1.0 files, the size of the Semantic Web would optimistically be increased by two orders of magnitudes. Although the study of Ding and Finin was of great importance as it was the first empirical study of the Semantic Web, this work has a number of limitations [16]. It's primary limitation was it was unknown if any of the Semantic Web documents indexed contained information that anyone would want to actually re-use. Intuitively, most of the data on this first-generation Semantic Web was likely to be of limited value. For example, the vast majority of data on the Semantic Web in 2006 was caused by Livejournal exporting every user's profile as FOAF – usually without the user's knowledge – without linking to other URIs, serving with the correct MIME type, and deploying `303` re-direction. The second main source of data in Ding and Finin's study, RSS 1.0, is also of limited value. RSS, originally an XML-based protocol generally used for newsfeeds, was given a RDF-compatible syntax, creating RSS 1.0 [6]. The very application of RDF in RSS 1.0 is questionable, as the data is primarily information about site updates, and so RSS 1.0 data is rarely merged, re-used, or even linked to in a manner that takes advantage of RDF. Due to the idiosyncratic nature of the data sources of the first generation Semantic Web, it is not surprising that the majority of the data likely contained little information that could *satisfy the information need* of the average user of the Web.

Due to the Linked Data initiative, the size of the Semantic Web has recently increased in size by several orders of

magnitudes due to the conversion of a large number of high-quality databases into RDF [12]. Since the study by Ding and Finin missed the rise of Linked Data, the time is ripe for more empirical studies of the Semantic Web. It is unclear how the dynamics of the Semantic Web are changing. While the number of URIs indexed by Linked Data search engines like Sindice shows that the general trend of the number of URIs on the Semantic Web visually follows a 'power-law,' the correct mathematical analysis has *not* been done to show this to be the case [26]. The only large-scale study of Linked Data at this time has been by Hausenblas et al., and it estimated the size of the Linked Data at approximately 2 billion triples [19]. The focus of that study was only on interlinking between data-sets, and it estimated that there were approximately 3 million interlinks between the various data-sets. The most popular interlinking property by far was *dbpedia:hasPhotoCollection*, with approximately 2 million occurrences, most likely to be due to the term being used by a Linked Data exporter around the popular photo-hosting service Flickr [2]. In summary, the Linked Data phenomenon is huge, much larger than the first-generation Semantic Web, and its properties have not been fully studied. In particular, there has been little work on determining how the issues of the reference of URIs play out in the wild given by Linked Data.

3. SAMPLING LINKED DATA VIA QUERY LOGS

The main problem facing any empirical analysis of the Semantic Web is one of *sampling*. As almost any database can easily be exported to RDF, any sample of the Semantic Web can be biased by the automated release of large, if ultimately useless, data-sets. This was demonstrated in an exemplary fashion by the release of RSS 1.0 data. RDF vocabulary terms that have little content, such as `rss:item`, quickly bias the statistical analysis. With the advent of Linked Data, this has to some extent already happened with large numbers of databases being released as Linked Data ranging from the BBC's John Peel recordings to the MusicBrainz audio CD collection [19]. How much of Linked Data is aimed for general use? Obviously, components like DBPedia, the export of Wikipedia to Linked Data, could be very useful [2]. The vast majority of data released into the Semantic Web is of appeal only to a niche audience, such as the large appeal of Bio2RDF to health care and life-sciences. Just as RSS 1.0 and the Livejournal export of FOAF biased sampling of the first-generation Semantic Web, the release of a large Linked Data set such as the Bio2RDF, containing approximately 65 million triples and so rivaling the size of DBPedia, can bias any sampling of Linked Data [7]. For example, if one just counted the number of URIs used on the Semantic Web, one would quickly find that `bio2rdf:xProteinLinks` would prove to be, in sheer number, a very popular term despite its relative lack of use outside the biomedical community. It is a small step then to imagine 'semantic spamming' that releases large amounts of bogus URIs into the Semantic Web. Furthermore, due to open nature of the Web, it is difficult, if not impossible, to determine how many actual separate providers of Semantic Web data there are, so a priori choosing seed samples or to 'weight' any sample is difficult. Unlike the original Web, which grew at least in an organic fashion for its first few years, the Web of Linked Data grows in very noticeable 'fits and starts' as large data-sets are released, so each data-set can vastly alter any empirical analysis. The question is not how to avoid bias in sampling, but *to choose the kind of bias one wants*. We are aiming for a bias towards the ordinary user of the Web.

What information is available on the Semantic Web that ordinary users are actually interested in, and how do we sample this data? The obvious candidate for exploring this would be look at a major search engine query log, as it gives a sample of the interests of many users in aggregate. Since Semantic Web search engines are currently used mostly by Semantic Web developers and not by ordinary users, the query log of a popular hypertext search engine should be sampled as opposed to a more specialized search engine. The entire bet of the Semantic Web is that it will contain information that many ordinary users will want to re-use and merge via Semantic-Web enabled applications, and that this information will primarily be about non-information resources such as entities like people and places and abstract concepts. Thus, the ideal sampling of the Semantic Web would be to extract query terms referring to physical entities and abstract concepts from a hypertext search engine query log, and then by virtue of a Semantic Web search engine we can determine precisely how much information Linked Data contains on these subjects.

3.1 The Live.com Query Log

There has been a much work in query log analysis in order to discover how to best satisfy the information needs of users on the Web. Since most search query logs of any size belong to search engines companies, it is often difficult for researchers outside those companies to analyze these query logs, and therefore most research in search query logs deal with small or special-purpose query logs, such as the Web track in the TREC competition [20]. A few employees of large search corporations have released detailed studies of their search engine query logs. In particular Silverstein et al.'s analysis of a billion queries in the Altavista query log is considered to be a large 'gold-standard' study of query logs [29]. In order to extract concepts and entities, we analyze the query log of approximately 15 million distinct queries from Microsoft Live Search, and all reference to the 'query log' are to this Microsoft query log, which is provided by Microsoft due to a 2007 'Beyond Search' award. This query log contains 14,921,285 queries. Of these queries, 7,095,302 (48%) were unique. Corrected for capitalization, 4,465,912 (30%) were unique. Of all queries, only 228,593 (2%) queries used some form of advanced keywords, while 709,102 (5%) used boolean operators and 266,308 (2%) used quotation, leading to a total of 1,204,003 (17%) queries using some advanced techniques provided by the search engines. The average number of terms per query was 1.76. Note that these extremely brief queries are normal for hypertext Web search engines, with an average query length of 2.35 being reported by Silverstein et al. for the Altavista query log [29]. Since we did not want to deal with queries that were only typed once or a few times, as these may not be representative of most user's interests, we did *not* select for further use any queries with a frequency less than 10, resulting in onlyfrom the total query log of 7,095,302, a reduction of 37%.

3.2 Extracting Queries for Entities and Concepts

Automatically classifying informational queries is difficult. Rule-based approaches that claim to work over entire query logs like those of Jansen et al. [21] are dubious at best, since they work by applying very loose specifications such as "query length greater than 2" and "any query using natural language terms." More promising work has applied both supervised and unsupervised machine-learning to discover informational queries, but only achieved an accuracy of 50% [3]. A number of machine-learning algorithms could be employed to learn named entities, but the sparse amount of linguistic context in query logs makes identifying a named entities difficult in a unsupervised manner, and there is virtually no labeled data for supervised learning [33]. Even most rule-based approaches for named entity recognition rely heavily upon capitalization and punctuation, such as 'I.B.M.' and 'Gustave Eiffel,' features that are lacking from query logs [23].

We call *queries that are automatically identified to be about physical entities in the query log* ***entity queries***. For the discovery of entity queries, people and places are obvious places to begin. An updated version of the system that was the highest performer at MUC-7 [23], a straightforward gazetteer-based and rule-based named entity recognizer, was employed to discover the names of people and places. The gazetteer for names was based on a list of names maintained by the Social Security Administration and the gazetteer for place names was based on the gazetteer provided by the Alexandria Digital Library Project. Although it could be possible to separate out people and places, this was not done. First, both of these are types of entities. Second, the names of many location such as 'Paris' or places like 'Georgia' can also be used as a name. This gazetteer-based approach was chosen to provide high precision, even at the cost of a dramatically reduced recall. This is an acceptable trade-off as we are attempting only to sample the number of queries that would likely to be have URIs on the Semantic Web. A high-quality sample of the query log is more important than a large one for this purpose. Of a random sample of 100 entity queries, a judge considered 94% to be correctly categorized as entities such as people or places.

From the pruned unique queries in the query log, totaling 4,465,912 queries, a total of 509,659 queries (11%) were identified as either people or places by the named-entity recognizer. The top 10 *entity queries* are given in Table 1. Some transactional and navigational queries, despite their relatively lower frequency overall in the query log, are highly clustered towards the top of the query distribution. These navigational queries such as 'chase' and 'office max' have clearly snuck into the top ten due to their use of common names in their website names. A legitimate number of real names, such as 'jessica alba' and 'marcus vick' were discovered.

A method for discovering abstract concepts in the query log is more challenging. These queries are called ***concept queries***, *queries that are automatically identified to be about abstract concepts in query log.* Previous attempts at discovering abstract concepts have employed machine-learning over truly massive query logs and document collections from Google [27]. Since this massive amount of data was not available, we employed WordNet instead. WordNet consists of approximately 207,000 words with unique synsets. Our algorithm for discovering abstract concepts in query logs using WordNet was straightforward: we only chose queries of length one where the query had a hyponym and hypernym, due to the difficulty of WordNet dealing with some multi-word queries. This assured that the query was for a class that was suitably abstract (having a hyponym) but not so abstract as to be virtually meaningless (had a hypernym). This resulted in a more restricted 16,698 concept queries (.4% of total query log). The top 10 concepts queries are given in Table 2. Again, a number of clearly transactional queries have managed to find themselves into the concept queries, such as 'chase' and 'drudge,' as well as a number of queries where the sense of a word has been taken over by a proper name, such as 'sprint' and 'aim.' Again, this is due to the preponderance of navigational names towards the top of the query distribution. Of a random sample of 100 concept queries, a judge considered 98% to be correct. The top ten concept queries are presented in Table 2. While some of the queries could be considered somewhat navigational (such as those for maps and dictionaries), they could all be considered informational queries about some abstract concept.

7311	david blaine
4039	kelly blue book
3053	chase
2997	jessica alba
2100	nick
1415	office max
1280	michael hayden
1139	harley davidson
1098	marcus vick
1092	keith urban

Table 1: Top 10 Entity Queries in Query Log

11383	weather
10321	dictionary
3675	people
3217	music
2192	autism
1468	map
1198	travel
1191	pregnancy
1104	news
1052	charter

Table 2: Top 10 Concept Queries in Query Log

3.3 Power-Law Detection

The frequency of queries, when rank-ordered, follows what is known as a 'power-law' distribution, with a relatively small number of very popular queries and a long-tail of queries only occurring once or twice, where most of the mass of the distribution is in the long tail and the 'top' of the distribution exponentially decreases. Since this distribution is common in search on the Web, we will define it precisely: A ***power-law*** is a relationship between two scalar quantities x and y of the form:

$$y = cx^{\alpha} + b \tag{1}$$

where α and c are constants characterizing the given power-law, and b being some constant or variable dependent on x that becomes constant asymptotically. Typically it is applied to rank-ordered frequency diagrams, where the frequency of some measurement is given on the horizontal axis while the rank order of the measurements in terms of their frequency is given on the vertical axis. The α exponent is the scaling exponent that determines the slope of the top of the distribution and provides the remarkable property of scale-invariance, such that if a true power-law is observed, as more samples are added to the distribution, the α remains constant, i.e. the distribution is 'scale-free' [32]. It is crucial to note that a power-law distribution violates assumptions of the normal Gaussian distribution, such that routine statistics such as averages and standard deviations can be and *usually* are misleading. In fact, one of the surest sign of a non-normal distribution like a power-law distribution is a very large standard deviation. Is such a distribution evident from Linked Data? One important question is how to detect power-law distributions in actual data. Equation 1 can also be written as:

$$\log y = \alpha \log x + \log c \tag{2}$$

When written in this form, a fundamental property of power-laws becomes apparent: When plotted in log-log space, power-laws are 'straight' lines. Thus,the most widely used method to check whether a distribution follows a power-law is to apply a logarithmic transformation, and then perform linear regression, estimating the slope of the function in logarithmic space to be α, as done by Ding and Finin [16]. However, standard least-square regression has been shown to produce systematic bias, in particular due to fluctuations of the long tail [14]. To determine a power-law accurately requires minimizing the bias in the value of the scaling exponent and the beginning of the long tail via maximum likelihood estimation. See Newman [25] and Clauset et al. [14] for the technical details.

Determining whether a particular distribution is a 'good fit' for a power-law is difficult, as most 'goodness-of-fit' tests employ normal Gaussian assumptions violated by potential power-law distributions. Luckily, the non-parametric Kolmogorov-Smirnov test can be employed for any distribution and so is thus ideal for use measuring 'goodness-of-fit' of a given finite distribution to a power-law function. While the details are given at length in Clauset et al. [14], intuitively the Kolmogorov-Smirnov test can be thought of as follows: Given a reference distribution P, such as an ideal power-law distribution generating function, and a sample distribution Q of size n suspected of being a power-law, where one is testing the null hypothesis that Q is drawn from P, then the Kolmogorov-Smirnov test compares the cumulative frequency of both P and Q to discover the greatest discrepancy (the D-statistic) between the two distributions. This D-statistic is then tested against the critical value p of the D-statistic at n, which varies per function. The null hypothesis is rejected if the D statistic is less than the critical p-value for n, p being the probability that the distribution was drawn from a power-law generating function given the estimated parameters. In order to determine how well the power-law method fits, whenever a power-law is reported, the D-statistic is also reported, and we will determine whether or not the fit was significant according to the conservative $p < .1$. The Kolmogorov-Smirnov test is valid even for power-law distributions since Q's cumulative density function is asymptotically normally distributed and this can be compared to the cumulative density function of P.

The query frequencies for entity and concept queries are plotted in logarithmic space in Figure 1. Both entity and concept queries appear to be linear in log-space, and so can be considered candidates for power-laws. Using the method described above, the α of the queries for entities was calculated to be 2.31, with long tail behavior starting around a frequency of 17 and a Kolmogorov-Smirnov D-statistic of .0241, indicating a significant good fit. The α of the queries for concept queries was calculated to be 2.12, with long tail behavior starting around a frequency of 36 with a Kolmogorov-Smirnov D-statistic of .0170, also indicating a significant good fit for the power law. Given their two remarkably similar α statistics and high goodness of fits, one can safely conclude that these query logs do indeed follow power-law distributions. This indicates our sample of entities and concepts are representative of the larger query log, which are well-known to follow power-law distributions [4].

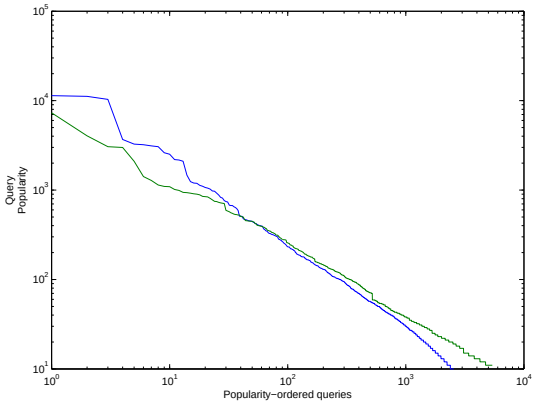

Figure 1: The rank-ordered frequency distribution of extracted entity and concept queries, with the entity queries given by green and the concept queries by blue.

3.4 Querying Linked Data with FALCON-S

Both the concept queries and the entity queries are used to query the Semantic Web. Since our goal was to discover how much of interest for ordinary users was present on the Semantic Web, one problem with using the entire query log was that it would contain a vast amount of unique queries that would likely to be never be repeated. So, we excluded a portion of the long tail from the study by removing all queries of less than a frequency of 10. The parameter 10 was chosen as it was the number that could reduce both entity and concept queries to the same order of magnitude. Due to the power-law behavior of both entity and concept queries, this truncation consists of 'removing' a large amount of the long tail, while maintaining the entire 'top' of the power-law distribution, as well as some significant component of the long tail. This procedure is justified insofar as the 'long-tail' likely consists of queries that are never or very rarely repeated, while the remaining queries represents queries that

are likely to be repeated. This pruning of low-frequency queries from our sampling does exclude many 'difficult' or 'specialist' queries, but we are aiming for queries that are general-purpose and popular. We call these *queries with more than 10 URIs returned from the Semantic Web* the ***crawled queries*** to distinguish them from the greater query log. Likewise, ***crawled entity queries*** are *entity queries with more than 10 URIs returned from the Semantic Web*, and similarly for ***crawled concept queries***.

This truncation reduced the amount of queries significantly, from 587,283 to 7,848 queries, removing 99% of the queries. It reduced the number of entity queries from 570,585 to 5,308 (a 91% reduction) and from the amount of concept queries from 16,698 to 2,540 (an 85% reduction). This gap in the result of pruning off the 'long tail' is interesting, as it shows that while there is a lower amount of concept queries than entity queries overall, concept queries are repeated by a order of magnitude or so more often than entity queries. The only caveat is that our identification of concept queries via WordNet is likely more stringent than our identification of entity queries, and thus leads to less concept queries overall. Furthermore, the vast majority of entity queries, as opposed to concept queries, appear to be queries that are only once or a very few times. This would make a certain amount of sense, as many queries for people and places are *not* for famous people and places, but for infrequently-mentioned people and places, such as `wayne way san mateo` and `sara matthews`. Some concepts that were as diverse as `gastropod` and `accolade`. Still, the crawled queries are still biased significantly in favor of entity queries, being composed of 68% being entity queries and only 32% concept queries.

The FALCON-S Object Semantic Web search engine [13] was used to query the Semantic Web for selected entity and concept queries between August 3rd and 4th 2008. We recognize that this a major weakness of the study, as its index may not be a representative sample of the entire Linked Data Web, but it is a significant sample regardless. At the time, FALCON-S seemed to have the best rankings, and a comparable index to other engines. The results of running the crawled queries against a Semantic Web search engine were surprisingly fruitful, although varying immensely. For entity queries, there was an average of 1,339 URIs (S.D. 8,000) returned per query. On the other hand, for concept queries, there were an average of 26,294 URIs (S.D. 14,1580) returned per query, with no queries returning zero documents. Given the high standard deviation of these results, it is likely that there is either a power-law in the resulting URIs for the queries, or some other non-normal distribution. As shown in Figure 2, when plotted in logarithmic space, both entity queries and concept queries show a distribution that is heavily skewed towards a very large number of high-frequency results, with a steep drop-off to almost zero results instead of the characteristic long tail of a power law. Far from having no information that might be relevant to ordinary user queries, the Semantic Web search engines returned either too many URIs possibly relevant to the query or none at all.

Another question is whether or not there is any correlation between the amount of URIs returned from the Semantic Web and the popularity of the query. As shown by Figure 3, there is *no* correlation between the amount of URIs returned from the Semantic Web and the popularity of the query. For entity queries, the Spearman's rank correlation statistic was

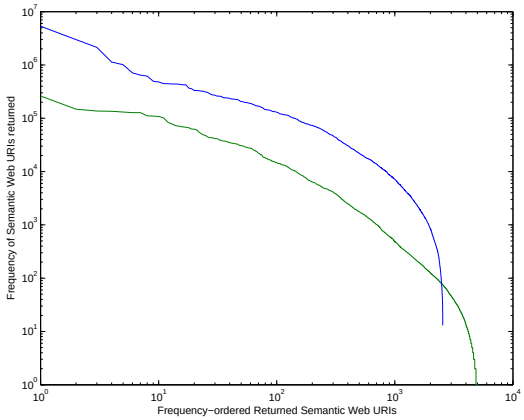

Figure 2: The rank-ordered frequency distribution of the number of URIs returned from entity and concept queries, with the entity queries given by green and the concept queries by blue.

the insignificant .0077 ($p > .05$), while for concept queries, the correlation was the still insignificant at .0125 ($p > .05$). Just because a query is popular or unpopular does not mean the Semantic Web will be more or less likely to satisfy the information need of the query. This makes sense, as the vast majority of queries are heavily dependent on current events and fashion, and the Linked Data data sources are not updated often enough to deal with this kind of information, so there is an inevitable temporal lag between the time information appears in the world outside the Semantic Web and its digitization on the Semantic Web. Yet as shown by Figure 2, the amount of *possibly* useful information for the vast majority of queries is still surprisingly large, although how many of the returned URIs are actually relevant to human users is not yet known.

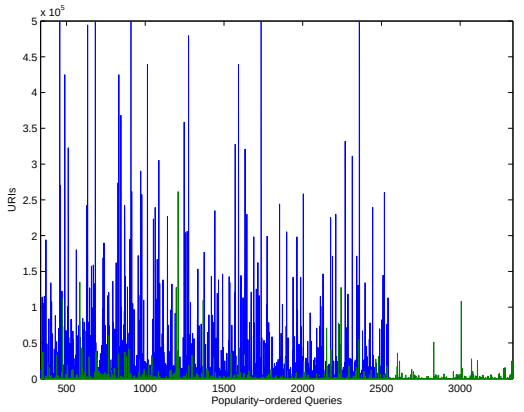

Figure 3: The rank-ordered popularity of entity and concept queries is on the x-axis, with the y axis displaying the number of Semantic Web URIs returned, with the entity queries given by green and the concept queries by blue.

4. EMPIRICAL ANALYSIS OF THE SEMANTIC WEB

Surprisingly, there is a deluge of possible Semantic Web URIs for any given query. Due to the high number of results for each query, we restricted our analysis to *the top 10 Semantic Web URI results for each query* as given by FALCON-S's ranking algorithm, and distinguish this subset from all the URIs returned by the Semantic Web, by calling these this subset the ***crawled URIs***. ***Concept URIs*** are *crawled URIs from the crawled concept queries* while ***entity URIs*** are *crawled URIs from the crawled entity queries*. Although crawled URIs are a small subset of the total URIs retrieved, given that user behavior in general inspects the first ten URIs returned by this search [18], it makes more sense to sample these ten URIs per query than to sample every URI retrieved. The crawled URIs totaled 70,128 URIs, composed of 25,400 (36%) concept URIs and 44,728 (63.78%) entity URIs. These URIs were crawled using HTTP GET with a preference for application-type of `application+rdf/xml` in order to prefer RDF files served by content negotiation, and any `303` redirection was followed.

Of all crawled queries, a total of 6,673 (85%) had at least 10 crawled URIs. All concept queries had at least 10 crawled URIs and only 4,133 of the entity queries (12% of all entity queries) did not have 10 queries. Inspecting just the set of queries that did not have 10 crawled URIs, the average number of URIs when 10 URIs were not returned were 2.89 (S.D. 2.88). So, the trend observed earlier was repeated in this smaller data-set, namely that while most of the time too many URIs are retrieved from the Semantic Web, sometimes there are *no* URIs are retrieved from the Semantic Web for certain entity queries. Looking at the data more closely, 357 (30%) of the crawled URIs with less than 10 results returned *no* URIs, while 138 (12%) returned a single URI and 113 returned two URIs (10%). These queries with zero results seem to be mostly for not well-known places such as `playa linda` (a hotel in Majorica) or fairly unknown people such as `william ravies` or misspellings or popular truncations of names for people such as `steven colbertbush`. This observation helps explains the sudden drop in Semantic Web URIs returned for queries in Figure 3. There was little overlap between the the crawled URIs retrieved by different queries, with an overlap in entity queries of 546 URIs (.01%) and an overlap in concept queries of 1031 URIs (.04%). In other words, the various queries weren't just retrieving the same small group of URIs over and over again.

4.1 URI-based Statistics

In this section, we inspect the various kinds of statistics we can detect on the 'macro-level' of the crawled URIs without actually accessing any Semantic Web documents from the URIs.

The HTTP status returned by attempting to access the various crawled URIs are given in Table 3. In particular, the most revealing statistic is the majority of the Semantic Web sampled by the crawled URIs is served using the `303` convention, not the hash convention. In fact, a total of 51,762 (73%) of crawled URIs use the 303 convention, while only 1,662 (2%) of the crawled URIs use the hash convention. Of these URIs returning the hash convention, manual inspection showed many to be FOAF files. This shows the vast majority of Linked Data is following the `303` convention and so obeying the W3C and the guide to publishing Linked Data [11]. This statistic as regards usage of the `303` convention is misleading in the broad sense, as most of the URIs are from a single source, DBPedia, as shown later in Table 4.

The majority of URIs, 51,873 (74%), served a Semantic Web document via `303` redirection, and so returned the `200` status code when the Semantic Web document was accessed after the redirection. `200` status codes without `303` redirection still form a substantial fraction of Semantic Web URIs. There are several reasons this; all hash convention URIs would by default still technically commit a redirect to be served by a `200` status code. However, this is only a minority (27%) of those URIs returning a `200` status code. The rest are likely caused by people serving RDF that does not have the access to the Web server configuration needed to serve RDF using the `303` redirection, while many others may have started serving RDF before the W3C TAG decision [28] was made or are not aware of Linked Data best practices. For example, some earlier RDF-enabled repositories like W3C WordNet did redirection by `300` redirection. A small percentage may be ordinary web-pages, perhaps containing some meta-data as enabled by GRDDL, that just happened to be indexed by the Semantic Web search engine [15]. Furthermore, of these crawled URIs, 9,156 (13%) URIs had no Semantic Web document that was accessible via HTTP, shown by the use of a `4xx` or a `5xx`-level status code.

51,873	73.97%	303
6,061	8.65%	200
4,517	6.44%	404
4,257	6.07%	500
3,147	4.49%	300
246	0.35%	406
20	0.03%	403
4	0.00%	302
3	0.00%	502

Table 3: Top 10 HTTP Status Codes for crawled URIs

The top 10 hosts of Semantic Web data in the crawled URIs is given by Table 4. DBPedia, the export of Wikipedia to RDF, dominates the results with 83% of all URIs coming from either Wikipedia or DBPedia [2]. The W3C itself is the third largest exporter of RDF with a share of 5%. Upon closer inspection, most of the URIs crawled from the W3C derive from the W3C-hosted export of the linguistic database Wordnet. The domain of the Freie Universität Berlin has a significant 2% of all RDF data, which is due primarily for its Flickr photo export to RDF. An RDF-version of Cyc and the biomedical data hosting site Bio2RDF also host small but significant amounts of Semantic Web data [22]. The Russian-blog hosting site `Liveinternet.ru` carries on the tradition of FOAF exporting of Livejournal. Truesense is another export of WordNet to RDF, although not as frequently used as W3C Wordnet. Towards the end of the ranking there is the RDF version of Univeristät Trier's widely used DBLP academic citation database and `Ontoworld.org`, a RDF-enabled wiki for the Semantic Web research community [31].

The average number of URIs hosted by a domain name

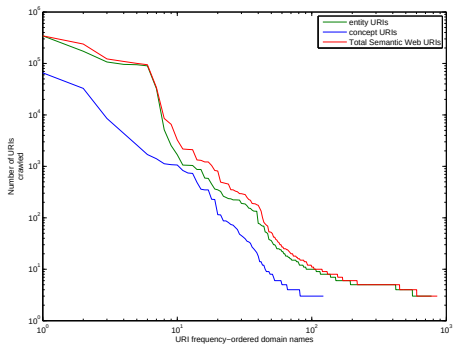

Figure 4: The rank-ordered distribution of the domain names hosting Semantic Web data from the crawled URIs ordered by number of URIs hosted.

was 1,268 (S.D. 16,060), with the average number of entity URIs hosted by any domain being 1,236 (S.D. 15,458) and the average number of concept URIs hosted by a domain being 1,0327 (S.D. 6,650). The very high standard deviations are usually a sign of power-law distribution, as shown in in Figure 4. Attempting to fit a power-law distribution, the α of the rank-ordered domain list frequency distribution is 1.53, with long tail behavior starting around 175 and a Kolmogorov-Smirnov D-statistic of .1414, indicating insignificant fit for the power-law distribution. In other words, while a few sources like DBPedia dominates the crawled URIs, with an rapidly decreasing number of smaller sites such as Cyc and the W3C, the long-tail individuals URIs hosting their FOAF files on their personal websites is still rather insignificant compared to the 'top' major sites hosting Linked Data. This is because the Linked Data is being artificially generated in large 'chunks' by projects like W3C Wordnet and DBpedia, and so do not organically form the power-law distribution characteristic of naturally-evolving complex systems.

54,698	78.00%	dbpedia.org
3,584	5.11%	wikipedia.org
3,448	4.92%	w3.org
1,704	2.43%	fuberlin.de
811	1.16%	cyc.com
701	1.00%	bio2rdf.org
599	0.85%	liveinternet.ru
417	0.59%	truesense.net
322	0.46%	dblp.unitrier.de
314	0.47%	ontoworld.org

Table 4: Top 10 Domain Names for URIs for Crawled URIs

4.2 Triple-based Statistics

In this section, we move our analysis down from the level of URIs to the level of the triples accessible from the URIs. Since a number of crawled URIs were inaccessible, this reduced the total number of ***accessible crawled URIs*** to 60,972, a reduction of (13%) from the crawled URIs. The accessible crawled URIs contained 24,074 accessible crawled concept URIs (95% of all crawled concept URIs) and 36,898 (82% of all crawled entity URIs) accessible crawled entity URIs. Thus, the accessible crawled URIs maintained a bias towards entity URIs (61% of all accessible crawled URIs) as compared to concept URIs (39% of all accessible crawled URIs). Each of the crawled accessible URIs was accessed, and this resulted in a total of 59,228 Web representations with only 48 URIs not allowing access to a Semantic Web document. These non-Semantic Web documents were usually ordinary web-pages from which RDF triples could not be extracted via GRDDL [15] or RDFa [1]. These crawled Semantic Web Documents we will call the ***crawled Semantic Web documents***, and the total sum of triples in these documents are called the ***crawled triples***.

There were a total of 411,574 RDF triples in the crawled triples, with 242,829 (59%) triples for concepts and 168,745 (41%) triples for entity URIs. Concepts, despite being fewer in number, seem to require more triples to describe than entities. The internal structure of these triples is of surprising interest. Of these triples, there were a total of 1,051 triples containing blank nodes, a measly .25% of all triples in the corpus, of which 772 (73%) were subjects and only 279 (27%) were in the object position. This means that the use of blank nodes, whose purpose is as syntactic placeholders in URIs for objects like lists and in representing n-ary arguments in RDF, is almost non-existent in our sample. Removing blank nodes, the composition was split between URI nodes (66%) and a surprisingly large minority of RDF literals nodes (34%). These literals contain some form of information in either 'unstructured' natural language or some form of structured information in a formal language, such as integer values.

Of the literals, a total of 403,119 were RDF string literals, while only 2% were of some other data type, with top 10 frequent data-types given in Table 5. The most frequent data-types are from XML Schema [10], while others are customized for DBPedia. It appears that the vast majority of RDF in the Semantic Web of interest to average users are simple URI-based triples with rich information in natural language. This also goes against the intuition that the vast majority of Semantic Web data that is of interest to ordinary users would be highly structured data of exported databases [8]. Instead, what is of interest in Linked Data is stored mainly in natural language, with RDF adding only a minimal structure to essentially fragments of natural language. While it could be argued that this particular finding is merely an artifact of DBpedia, however, it should be acknowledged that DBpedia is, given that our querying includes other data-sets, this finding may well be generalizable. We are not studying the Semantic Web as some of its designers would *like* to have it, but as it actually exists, and part of its existence is that DBpedia forms a huge central cluster that for ordinary users is the most interesting and useful part of Linked Data.

One interesting question is the predominance of the various kinds of Semantic Web knowledge representations terms on the Semantic Web, since this would show what kinds of inference could actually be deployed on the Semantic Web. First, of the total 1,093,212 URIs in triples harvested from the crawled accessible URIs, only 243,776 (22%) were from one of the primary W3C Semantic Web knowledge representation languages, either RDF, RDF(S), or OWL.

403,119	97.95%	RDF plain literal
3,103	0.75%	w3c:/XMLSchema#integer
2,789	0.68%	w3c:/XMLSchema#string
1,185	0.29%	w3c:/XMLSchema#double
522	0.13%	w3c:/XMLSchema#date
248	0.06%	w3c:/XMLSchema#float
136	0.03%	w3c:/XMLSchema#gYear
65	0.02%	w3c:/XMLSchema#gYearMonth
59	0.01%	dbpedia:Rank
46	0.01%	dbpedia:Dollar
14	0.00%	w3c:/XMLSchema#int
9	0.00%	dbpedia:Percent

Table 5: Common Data Types in Crawled Triples

Of these, the RDF vocabulary itself was the most popular, with 109,300 URIs (45%), followed fairly closely by the RDF(S) vocabulary with 100,340 URIs (41%), and OWL being dwarfed by RDF and RDF(S) with only 34,136 URIs (14%). This does not mean that OWL is irrelevant to the other corpus, as ontologies constructed with OWL could be deployed to model the concepts and entities employed in 'instance' data. Yet while OWL has been an academic success story, insofar as practical deployment, RDF terms and RDF(S)-based inference seems to be the foundation of the Semantic Web in practice.

What precise URI-based terms are used in these knowledge representation languages? The top constructs in either RDF, RDF(S), or OWL in crawled triples are given in Table 6. To summarize, RDF(S) class and sub-class reasoning is very popular, with this construction consisting of nearly half (48%) of knowledge representation use of the Semantic Web. The second most popular use of knowledge representation (22%) is for natural language annotation, describing a particular Semantic Web resource using natural language and connecting this natural language description to the URI via the use of `rdfs:comment` or `rdfs:label`. There are surprisingly few (4%) actual ontologies in the crawled Semantic Web resources. Furthermore, non-traditional features of RDF(S), such as the use of `rdfs:property`, are frequent occurrences. Even reification of RDF triples, officially discouraged by the Semantic Web community, accounts for only 95 triples, and there is also fairly heavy use of discouraged RDF constructs to represent different kinds of lists, such as `rdf:Alt` (349 occurrences) and `rdf:Bag` (344 occurrences). Lastly, while many Semantic Web researchers originally hoped that the use of inverse functional properties would allow the merger of Semantic Web data, there were zero explicitly declared usages of `owl:inverseFunctionalProperty`. Overall, the usage of OWL, RDF(S), and RDF terms in the corpus also follows to some degree a power-law like distribution, where α equal to 1.5, with long tail behavior starting around 90, although the Kolmogorov-Smirnov D-statistic of .1911 reveals this to insignificant. This is because while a few terms vastly dominate, the vast majority of other terms are *not used at all*. This has reprecussions for both Semantic Web implementers and vocabulary specification within the W3C, since obviously some level of concentration of effort upon the most frequently-deployed terms would be reasonable.

One of the most popular OWL constructs is indeed the controversial `owl:sameAs` term, which is used to declare some sort of global equivalence between two URIs. While a tiny portion (.47%) of overall Semantic Web modelling term usage, it is far from insignificant, with 1,157 occurrences. The use of `owl:sameAs` in the wild is far different than the role it plays in popular debate within the Semantic Web community would suppose. Logicians hold that `owl:sameAs` is only for what is properly considered individuals in description logic, so that classes and properties should use the more restricted and semantically correct `owl:equivalentClass` and `owl:equivalentProperty`. Yet this best practice in logic hasn't the Linked Data community, as `owl:equivalentClass` has only 2 occurrences and there are none of `owl:equivalentClass`. Instead, the Linked Data movement uses `owl:sameAs` to simply "state that another data source also provides information about a specific non-information resource," so leading `owl:sameAs` to tend to mean 'more-or-less the same thing as' [11]. This practice leads to the fear that the use of `owl:sameAs` would propagate too far, such that many URIs for the perhaps differing referents would be declared identical [17].

Both critiques of `owl:sameAs` appear to be wrong. Given the amount of Semantic Web URIs returned by the queries, while there is considerable use of `owl:sameAs`, it appears that the manual discovery and publication of co-referential URIs using `owl:sameAs` falls far behind the actual growth of Linked Data. One could say that `owl:sameAs` is not being used enough. The real problem is not that distinct things are being given the same URI, but the *reverse*; namely that it appears endemic that the same thing has multiple URIs. So Berners-Lee's hypothesis appears to be wrong: A single thing is likely identified by more than a single URI on the Semantic Web.

73,451	30.31%	rdfs:Class
47,044	19.30%	rdfs:comment
44,113	18.10%	rdfs:subClassOf
8,630	3.54%	owl:Ontology
7,256	2.97%	rdfs:label
6,618	2.14%	rdf:Subject
5,107	2.09%	owl:ObjectProperty
3,642	1.49%	rdfs:subPropertyOf
1,157	0.47%	owl:sameAs
535	0.29%	rdfs:range

Table 6: RDF and OWL Constructs in Crawled Triples

The top 10 Semantic Web vocabularies used in the crawled triples, including those outside of the W3C-approved Semantic Web knowledge representation languages, are shown in Table 7. The results should not be that surprising, in particular the vast dominance of DBPedia. Perhaps surprising is the surprising amount of usage of Cyc terms, as well as terms from SKOS, the Simple Knowledge Organization System of the W3C, whose primary source of deployment is the W3C's export of WordNet to RDF [24]. FOAF is also significant, although not nearly as dominant as was found earlier by Ding and Finin [16]. Also popular is YAGO (Yet Another Global Ontology), a merger of WordNet and Wikipedia category hierarchies employed by DBPedia [30].

366,849	33.55%	DBpedia URIs
109,300	9.99%	RDF URIs
100,340	9.17%	RDF(S) URIs
94,520	8.65%	Cyc URIs
34,136	3.12%	OWL URIs
6,563	0.60%	SKOS URIs
4,728	0.43%	dblp.l3s.de
3,263	0.29%	FOAF URIS
2,170	0.20%	YAGO URIs
1,836	0.16%	WordNet URI

Table 7: Top Vocabulary URIs in Crawled Triples

5. CONCLUSION

The empirical analysis of Linked Data presented in this study is by no means complete, for it is only a moderately small sample by one Semantic Web search engine (and so hurt or benefit by the idiosyncratic behavior of the searching of FALCON-S), although it is an important one as this sample is driven by Web search queries by actual users. The results of this empirical analysis show a transformation from the first-generation Semantic Web to the next generation Web of Linked Data. The Semantic Web as it existed in the first-generation was a motley collection of RDF triples, heavily dominated by a few exports of social networking data into FOAF and a long-tail of complex academically-produced ontologies. Linked Data - at least the section of it that is of interest to users querying the Web for information - is dominated heavily by DBPedia and consists primarily of collections of triples that provide a minimal structure to natural language [16].

On the level of triples, there are some surprising conclusions. The triples on the Semantic Web contain a vast range of data, and the exact kinds of URIs used in the triples are somewhat unpredictable. However, the kinds of vocabularies actually deployed are almost entirely from a few large vocabularies, such as DBPedia, DBLP, WordNet, YAGO, and FOAF. This again points to a victory of Berner-Lee's idea that a few large vocabularies with well-defined terms could dominate the Semantic Web [9]. The kinds of triples that structured this data do not contain many OWL terms optimized for inference, but consist almost entirely relatively straight-forward RDF(S) expressions for sub-class relationships and for annotations in natural language. Overall, Linked Data is primarily being used to provide structured relationships between fragments of natural language, and *not* for inference.

One could argue that that these results are more characteristic of FALCON-S and DBpedia than the second-generation Linked Data as a whole. However, we would respond that it is natural in decentralized information systems for power law distributions, where one source of data massively outweighs others in weight to evolve, and the 'giant component' of Linked Data is DBpedia [5]. In fact, if such a 'giant component' and long tail were not observed, it would be cause for suspicion. In conclusion, there is potentially lots of rich information that ordinary Web search users in Linked Data form, and so one outcome of this analysis should be a greater interest in Linked Data from even mainstream information retrieval systems. However, for future work we wish to repeat this study over different Semantic Web search engines beside FALCON-S, which we recognize is a major limiting factor. Second, there is likely too many URIs in Linked Data for a given query, although to truly substantiate this claim ideally the URIs returned by the search engines should each be individually inspected, although this is difficult in practice. Yet even at this point it seems is likely that there are many co-referential URIs for the 'same thing' that are not explicitly modelled with `owl:sameAs`, and unless action is taken this growth of URIs will contine of the future. Unless there is URI re-usage many of the data-sources for Linked Data are more like semantic islands rather than parts of interconnected semantic continents.

6. ACKNOWLEDGEMENTS

Harry Halpin was supported in part by a Microsoft "Beyond Search" grant.

7. REFERENCES

[1] B. Adida, M. Birbeck, S. McCarron, and S. Pemberton. RDFa in XHTML: Syntax and Processing. W3C Recommendation, W3C, 2008. http://www.w3.org/TR/rdfa-syntax/.

[2] S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the International and Asian Semantic Web Conference (ISWC/ASWC2007)*, pages 718–728, Busan, Korea, 2007.

[3] R. Baeza-Yates, L. Calderon-Benavides, and C. Gonzalez. Understanding user goals in web search. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, pages 98–109, 2006.

[4] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley-Longman, New York City, New York, USA, 1999.

[5] A.-L. Barabasi, R. Albert, H. Jeong, and G. Bianconi. Power-law distribution of the World Wide Web. *Science*, 287:2115, 2000.

[6] G. Beged-Dov, D. Brickley, R. Dornfest, I. Davis, L. Dodds, J. Eisenzopf, D. Galbraith, R. Guha, K. MacLeod, E. Miller, A. Swartz, and E. van der Vlist. RDF Site Summary (RSS) 1.0. Technical report, http://web.resource.org/rss/1.0/spec, 2001.

[7] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706–716, 2008.

[8] T. Berners-Lee. What the Semantic Web can represent, 1998. Informal Draft. http://www.w3.org/DesignIssues/rdfnot.html (Last accessed on Sept. 12th 2008).

[9] T. Berners-Lee and L. Kagal. The fractal nature of the Semantic Web. *AI Magazine*, 29(3), 2004.

[10] P. Biron and A. Malhotra. XML Schema Part 2: Datatypes. Recommendation, W3C, 2004. http://www.w3.org/TR/xmlschema-2/.

[11] C. Bizer, R. Cygniak, and T. Heath. How to publish Linked Data on the Web, 2007. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ (Last accessed on May 28th 2008).

[12] C. Bizer and A. Seaborne. D2RQ: Treating non-RDF databases as virtual RDF graphs. In *Proceedings of International Semantic Web Conference*, 2004.

[13] G. Cheng, W. Ge, and Y. Qu. FALCONS: Searching and browsing entities on the semantic web. In *Proceedings of the the World Wide Web Conference*, 2008.

[14] A. Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data, 2007. http://arxiv.org/abs/0706.1062v1 (Last accessed October 13th 2008).

[15] D. Connolly. Gleaning Resource Descriptions from Dialects of Languages (GRDDL). Technical report, W3C, 2007. Recommendation.

[16] L. Ding and T. Finin. Characterizing the Semantic Web on the Web. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 242–257, 2006.

[17] A. Ginsberg. The big schema of things. In *Proceedings of Identity, Reference, and the Web Workshop at the WWW Conference*, 2006. http://www.ibiblio.org/hhalpin/irw2006/aginsberg2006.pdf.

[18] L. Granka, T. Joachims, and G. Gay. Eye-tracking analysis of user behavior in www search. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 478–479, New York, NY, USA, 2004. ACM.

[19] M. Hausenblas, W. Halb, Y. Raimond, and T. Heath. What is the size of the Semantic Web? In *Proceedings of Conference on Semantic Systems (iSemantics)*, Graz, Austria, 2008. http://tomheath.com/papers/hausenblas-isemantics2008-size-of-semantic-web.pdf.

[20] D. Hawking, E. Voorhees, N. Craswell, and P. Bailey. Overview of the trec-8 web track. In *Proceedings of the Text REtrieval Conference (TREC)*, pages 131–150. ACM, 2000.

[21] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Information Process and Management*, 44(3):1251–1266, 2008.

[22] D. Lenat. Cyc: Towards programs with common sense. *Communications of the ACM*, 8(33):30–49, 1990.

[23] A. Mikheev, C. Grover, and M. Moens. Description of the LTG system used for MUC. In *Seventh Message Understanding Conference: Proceedings of a Conference*, 1998.

[24] A. Miles and S. Bechhofer. SKOS Simple Knowledge Organization System reference. Working draft, W3C, 2008. http://www.w3.org/TR/skos-reference/.

[25] M. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323–351, 2005.

[26] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: a document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics, and Ontologies 2008*, 3(1):37–52, 2008.

[27] M. Paşca. Weakly-supervised discovery of named entities using web search queries. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (CIKM)*, pages 683–690, New York, NY, USA, 2007. ACM.

[28] L. Sauermann and R. Cygniak. Cool URIs for the Semantic Web. Technical report, W3C Semantic Web Interest Group Note, 2008. http://www.w3.org/TR/cooluris/.

[29] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[30] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: a core of semantic knowledge. In *In Proceedings of the 16th International Conference on World Wide Web*, pages 697–706, New York, NY, USA, 2007. ACM.

[31] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, and R. Studer. Semantic wikipedia. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 585–594, New York, NY, USA, 2006. ACM.

[32] D. Watts and S. Strogatz. A review of ontology based query expansion. *Nature*, 6684(393):409–410, 1998.

[33] C. Whitelaw, A. Kehlenbeck, N. Petrovic, and L. H. Ungar. Web-scale named entity recognition. In *Proceedings of Conference on Information and Knowledge Management*, pages 123–132. ACM, 2008.