Predicting Website Correctness from Consensus Analysis

Steven O’Hara
University of Texas San Antonio
One UTSA Circle
San Antonio, TX 78249
(1) 210-458-4436
sohara@cs.utsa.edu

Tom Bylander
University of Texas San Antonio
One UTSA Circle
San Antonio, TX 78249
(1) 210-458-5693
bylander@cs.utsa.edu

ABSTRACT

Websites vary in terms of reliability. One could assume that NASA’s website will be very accurate for Astronomy questions. Wikipedia is less accurate but is still more accurate than a generic Google search. In this research we ask a large number of “factoid” questions to several different search engines. We collect those responses and determine the correctness of each candidate answer. The answers are grouped by website source, and are compared to other websites to infer website correctness.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *question-answering (fact retrieval) systems.*

General Terms

Algorithms, Reliability, Verification.

Keywords

Text Retrieval, Question Answering, Intelligent Search, Answer Consolidation, Website Reliability.

1. INTRODUCTION

This paper is focused on “factoid” type questions [1], such as “What is the area of Vietnam?” or “What is the atomic mass of Copper?”. These questions can be characterized as:

```
<domain, class, object, attribute>
```

tuples, where the result is a consolidated answer of the form:

```
<value, delta, units, source, confidence>
```

The preceding two questions are represented by:

```
<Geography, country, Vietnam, area>
<Chemistry, element, Copper, atomic mass>
```

with possible answers:

<331210, 5, km^2, Wikipedia, 0.8>
<63.546, 0.0005, amu, Wikipedia, 0.99>

In many cases, you may already have a website in mind, such as NASA (www.nasa.gov), Wikipedia (www.wikipedia.org), the CIA World Factbook (www.cia.gov/library/publications/the-world-factbook) or IMDb (www.imdb.com). You would expect these sites to be more reliable than a generic search.

Search websites use various algorithms to rank-order candidate answer websites. Google, for example, uses a page ranking algorithm. Their assumption is that the more websites reference a particular site, the more reliable it is [2].

Because the number of websites is steadily increasing, it is easy to be overwhelmed with candidate answers, without any definitive way to know which answers are the best.

This paper introduces a methodology to obtain a consensus answer for a given factoid question. By broadcasting a large number of questions, and assessing their consensus answers among websites and the correctness of these answers, we show that websites that agree with consensus answers are more likely to provide correct answers.

2. PRIOR RESEARCH

2.1 TruthFinder

The TruthFinder system is heavily referenced by other relevant literature in this area.

Yin [3] discusses what they call the “veracity” or “conformity to truth” of web sites. They develop an algorithm called TruthFinder for identifying the correct answer among conflicting information.

Their algorithm is based on the idea that a fact is likely to be true if it is provided by trustworthy web sites, and a web site is trustworthy if most of the facts it provides are true. They use an iterative approach where both the probabilities of the facts being true and the trustworthiness of the web sites are inferred from each other.

They show that this approach outperforms both a simple voting scheme and the ranking returned by a generic Google search.

Our research extends the TruthFinder idea by showing that consensus answers are more likely to be true; in particular, that websites that tend to provide an answer that agrees with the consensus answer are more likely to provide the correct answer for new questions.

2.2 Integrating Conflicting Data

Dong [4] looks at the problem of finding “true” values on the internet. A simple voting technique will often fail because web sites will often copy information from another web site, yielding false votes.

They use an iterative Bayesian approach to build a source-dependency graph between web sites. An assumption is made that there is one “true” value and many distinct “false” values for

RACS’12, October 23-26, 2012, San Antonio, TX, USA.
Copyright 2012 ACM 978-1-4503-1492-3/12/10…$15.00.

particular objects. Using this assumption, the sharing of “false” values is typically a rare event when the sources are fully independent.

Our research is complementary to this approach as our research allows slightly different candidate answers to support each other.

2.3 Corroborating Information

Galland [5] starts with a list of reasons why answers from the internet are often wrong:

- Disagreement: e.g., the birth date of Napoleon Bonaparte, which determines whether Napoleon was born French or Italian, is reported as August 15, 1769 or January 7, 1768.
- Outdated sources: information changes over time, such as where somebody works.
- Simple errors: misspellings, name changes, typographic errors, estimations, etc.

They develop several different algorithms that estimate the truth values of facts and the trust of sources. They also use TruthFinder as a basis for comparison. Each of these algorithms is shown to outperform simple voting.

The central theme of their research, namely that overall trust-worthiness of a website is correlated with accuracy of individual facts within that website, is a key concept for our research effort.

3. METHODOLOGY

3.1 Ontology

One technique used to reduce irrelevant websites from the search is to introduce a high-level ontology. In our <domain, class, object, attribute> tuple, all objects belong to a domain, such as Astronomy, Geography or Chemistry. Additionally, objects are members of a class hierarchy in that ontology. So, Mars and Jupiter are both instances of Planets, all of which are Celestial Objects. This knowledge in turn determines what attributes are applicable for each object in a given class and domain. Planets have attributes like Diameter and Mass; Elements have an Atomic Number and Atomic Mass; Countries have Capitals, Populations and Areas.

Given this ontology, it is possible to determine many objects of each class in the domain as well as many attributes for each object. Given *M* objects with *N* attributes, it is possible to generate $M \times N$ factoid questions. These questions can be sent to a wide range of websites, as well as many search engines like Google, Bing, Wolfram Alpha, etc.

Our previous work [6], which has been effective in rank ordering multiple answers, in terms of likelihood of correctness.

3.2 Consolidating Candidate Answers

The key concept for consolidating consensus answers is the measure of Support, which we define as a quantifiable measure of how much one candidate answer “supports” another candidate answer. Identical answers will have a measure of 1.0. Disjoint answers will have a measure of 0.0. Overlapping answers will have a measure in between.

When the attribute is a number, we use implied precision (along with unit conversion) in determining how much one candidate answer supports another. Dates are treated in a similar manner in which the year, month, and day (or lack of month and day) are used to determine an implied precision. When the attribute is a string, the support is based on character and word matching.

3.2.1 Numeric Questions

Numeric query answering is different than ordinary query answering [7] in two fundamental ways: implied precision and unit conversions. Query accuracy can be improved by taking advantage of these distinct differences.

In a few cases, numbers have explicit precision, for example, a value in the form “2100 ± 40 grams”. However, most resources will report values like “2100 grams” or “2110 grams” etc. Based on common interpretation of numbers ending in zeros, we assume that the 2100 value has an implied precision of 2050 to 2150, and that the 2110 value means 2105 to 2115. Likewise, the value 2050 is assumed to mean 2025 to 2075. These interval values can be compared against each other to see how much they reinforce each other.

Given two numeric answers X and Y, already converted to the same units, and each with its own positive delta interval ($\Delta X > 0$ and $\Delta Y > 0$), define their ranges:

$$X_{LO} = X-\Delta X \quad X_{HI} = X+\Delta X \quad Y_{LO} = Y-\Delta Y \quad Y_{HI} = Y+\Delta Y$$

To determine the amount of Support X gives to Y, use:

$$(\min(X_{HI}, Y_{HI}) - \max(X_{LO}, Y_{LO})) / (X_{HI} - X_{LO})$$

For example, suppose A ranges from 10 to 60 while B ranges from 40 to 140, then Support is calculated as:

$$\text{Support}(A,B) = (60-40) / (60-10) = 40\%$$

$$\text{Support}(B,A) = (60-40) / (140-40) = 20\%$$

So 40% of A’s interval is in B, and 20% of B’s interval is in A.

3.2.2 Date / Time Questions

When the attribute is expected to be a date (“When was Lincoln born?”), similar reasoning can be applied to obtain an implied precision. A candidate answer of 1809 supports an answer of February 1809, which in turn supports the answer of February 12, 1809. But March 1809 does not support February 1809. And a year like 1810 is not assumed to have an implied interval of five years.

Table 1. Date Support Computations

Year Condition	Month Condition	Day Condition	Support
$X_{YR} \neq Y_{YR}$	-	-	0.0
$X_{YR} = Y_{YR}$	Y_{MON} missing	-	1.0
“	X_{MON} missing	-	1 / 12
〃	$Y_{MON} \neq X_{MON}$	-	0.0
“	$Y_{MON} = X_{MON}$	Y_{DAY} missing	1.0
“	“	X_{DAY} missing	1 / 30
〃	〃	$Y_{DAY} \neq X_{DAY}$	0.0
“	“	$Y_{DAY} = X_{DAY}$	1.0

Each candidate answer for a date is considered to be a year, a year-month or a year-month-day. For this effort, we are not considering decades or centuries. Given two candidate answers X and Y, each with optional month and day, compute Support using Table 1.

According to this computation, a date like Feb 12, 1809 has zero support for Feb 13, 1809. This is equivalent to a number like 12.0 ± 0.5 which has zero support for 13.0 ± 0.5.

3.2.3 Textual Questions

For textual attributes ("What is the capital of Australia?"), incorrectly spelled answers can still support the correct answer, so there is still an opportunity for different answers to support each other. Canberra is the capital, but answers like Canbera and Canbeera are both supportive.

In this research, Levenshtein's edit distance [8] is used to compare two strings for mutual support. It is basically the number of single character edit changes (insert, delete or substitute) needed to map from one string to the other. For example, the edit distance between "cat" and "canoe" is three (one substitution and two insertions), and the edit distance between "Boise" and "Bossy" is two (both substitutions).

Normalizing these support values to be in the range 0 to 1 is accomplished by dividing the edit distance by the length of the longer string, and subtracting it from 1. For strings S and T:

$$\text{Support}(S, T) = 1 - \text{EditDistance}(S, T) / \text{Max}(S_{LEN}, T_{LEN})$$

3.3 How to Select the "Best" Answer

Given a set of candidate answers, and a support matrix of how they support each other, it is easy to determine which candidate answers receive the most support, and which candidate answers give the most support. It appears reasonable to use either method to determine the "best" answer, but there is a big difference.

Choosing the answer that receives the most support will tend to get a very broad answer that spans many other answers.

A more general answer, e.g., 3.14 for π, will tend to be supported by more specific answers, e.g., 3.14159. Our assumption is that many answers tend to be approximations of the correct/accepted answer. In order to prefer more specific answers over more general answers, we rank candidate answers by the amount of support they give to other candidate answers, rather than the amount of support they receive.

For example, consider the question, "What is the diameter of Pluto?" Five of the 25 candidate answers were over 1,500 miles and are not shown in Figure 1. The remaining candidate answers are sorted by their interval size.

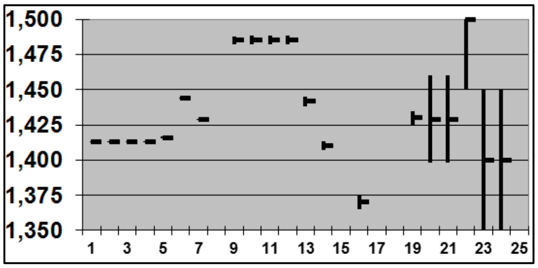

Figure 1. Diameter of Pluto in Miles

It is easy to see that there are a variety of answers, and there is also a wide range of implied interval sizes. The two candidate answers of 1,400 miles receive the most support from other candidate answers. The four candidate answers of 2,274 kilometers (1,413 miles) give the most support to the other candidate answers.

Our ranking prefers the specific answer of 1,413 miles to the general answer of 1,400 ± 50 miles in the above example. According to NASA, the mean diameter of Pluto is 2,302 km, which is about 1,430 miles.

3.4 Rightiness

We define a measure called "rightiness" that is a measure between zero and one for how correct a candidate answer is, relative to a known target answer. We call it "rightiness" (a deliberate pun to "truthiness") both because the target answer is not necessarily right and because we do not claim our measure is correct by some outside standard. Nevertheless, our design of the measure is intended so that differences between rightiness values roughly correspond to a similar change in accuracy.

After significant preliminary analysis, we derived these measures of rightiness empirically. It was necessary to determine how close candidate answers were to the target answers, taking into account the fact that all answers have intervals, whether implied or explicit. None of the previous research we mentioned addresses this issue.

It is computed differently for each of the three types. For each of these, we define:

C is the candidate answer, and
T is the target answer.

Rightiness for numbers roughly measures how many significant digits match, up to 3 digits. It will result in about 0.3 if only the first significant digit matches, about 0.6 if the first two significant digits match, and about 0.9 if the first three significant digits match. A result of 1.0 is an exact match. Assuming C and T are positive,

$$R(C, T) = 1 - \min(C, T) / \max(C, T)$$
$$\text{Rightiness}(C, T) = \max(0, -\log_{10}(R(C, T) + 0.001) / 3)$$

For numbers, accuracy can be roughly measured by the number of matching significant digits, but target answers can vary in the number of significant digits. To ensure that our measure is comparable between different questions, we decided to emphasize the first three significant digits. So rightiness for numbers roughly measures how many significant digits match up to three digits.

Rightiness for dates is based on how far apart the two candidate answers are, with a maximum of 1.0 for two dates that are identical, and a minimum of 0.0. The subscript Y means year, M means month, and D means day.

$$R(C, T) = \left| 365*(C_Y - T_Y) + 30*(C_M - T_M) + (C_D - T_D) \right|$$
$$\text{Rightiness}(C, T) = 1 - \min(1, R(C, T) / 365)$$

Rightiness for strings uses Levenshtein's edit distance, which is the minimum number of character additions, deletions, and substitutions to derive one string from the other.

$$R(C, T) = \text{EditDistance}(C, T)$$
$$\text{Rightiness}(C, T) = 1 - R(C, T) / \max(C_{LEN}, T_{LEN})$$

Note that our measure of support does not correspond to rightiness. For example, the support between 3.14159 and 3.14195 is zero in both directions, but the support between 3.14159 and 3 is nonzero (note that the implied precision of 3 suggests a range of 2.5 to 3.5). However, if the target answer is 3.14159, a candidate answer of 3.14195 is certainly more accurate than 3.

4. EXPERIMENTAL RESULTS

In this section, we describe how we set up the experiments, work through a few sample Astronomy questions, and show results. Then we break down the results along several axes to show that the results are consistent across several dimensions.

4.1 Algorithm

The following processing steps are used to determine website relative reliability.

4.1.1 *Collect Questions*

An offline process is used to collect questions. For this research, we have been focusing on finding questions with many candidate answers. We have 2,945 factoid questions summarized in Table 2.

Table 2. Experimental Questions

Domain (Classes)	Questions
Astronomy (Planets)	47 numeric
Chemistry (Elements, Molecules)	1,468 numeric
Geography (Lakes, Capitals, etc.)	867 numeric, 289 string
History (President's Birthdays)	44 date
Sports (Baseball Players)	230 date

4.1.2 *Training and Test Sets*

Half of the questions were used during the training / development phase. The analysis and results presented in this paper are based on the other half of the questions, none of which were used during the training phase.

4.1.3 *Collect Answers from Experts*

We use the term "expert" to indicate that a software wrapper has been written for that website. It could be for a search engine like Google or Bing, or for a specific website, like NASA. This differs from a website source. We use the term "source" for the domain name part of the actual website URL that contains the answer. For example, Google itself is almost never a source, the websites that it references are the sources.

Several different experts are used for answering questions. Google, Bing, Wolfram Alpha, Ask, NASA, CIA and FACTO are each sent all questions. An average of approximately 34 candidate answers were returned for each question, ranging from two to several hundred.

Our experts for search engines, such as Google or Bing, present candidate answers from a wide range of sources. Our NASA expert, on the other hand, refers to itself as the source.

4.1.4 *Calculate Scaled Support and Rightiness*

As described earlier, support values are calculated for each candidate answer for each question. This yields a measure of how much support each candidate answer gives other candidate answers for the same question.

Raw support values are always zero or positive, but they will generally be larger for a question with more candidate answers. Hence, a normalization is done to map the range to the interval [0..1], by dividing by the largest raw support value. These are denoted by the term "scaled support". In many cases, a single website will have candidate answers to multiple questions, in which case the term "average scaled support" is used.

Each question has a target answer. All 955 molecular target answers came from www.reciprocalnet.org; all 230 baseball player target answers came from www.Baseball-reference.com; all other target answers came from Wikipedia.

The rightiness is calculated for each candidate answer by comparing it to the target answer for each question.

4.1.5 *Collect Statistics by Website Domain*

Every candidate answer has a website source. Considering just the domain part of the website, we collect all the answers by website domain, excluding the websites where we obtained the target answers. Any website with less than ten candidate answers is omitted from our analysis, except for Figure 6.

4.1.6 *Correlate Scaled Support to Rightiness*

The hypothesis of this paper is that our measure of Scaled Support is highly correlated with the measure of Rightiness. In other words, the higher the Scaled Support level for a range of questions (within a domain), the more reliable that website will be.

In the next section, we show experimental results for this algorithm.

4.2 Sample Analysis

In this section, we work through a few sample questions to clarify the algorithm expressed in the previous section.

4.2.1 *Sample Questions and Responses*

The questions shown in Table 3 were chosen because they are all typical questions from the same domain, Astronomy. Our hypothesis is that websites that perform well on these questions will also perform well on other Astronomy questions.

The NASA website (nssdc.gsfc.nasa.gov) had candidate answers to these questions. Target answers came from en.wikipedia.org.

Table 3. Astronomy Target Answers

Q	Answers	Object	Attribute	Target
44	107	Venus	Diameter	12105.6 ± 1.0 km
2	95	Earth	Diameter	12756.2 ± 1.0 km
14	104	Mars	Diameter	6786.3 ± 1.0 km
10	45	Jupiter	Radius	69911.0 ± 6.0 km
36	50	Saturn	Radius	60268.0 ± 4.0 km
26	45	Neptune	Radius	24764.0 ± 15.0 km

For two of the questions, the Google search answer also provided answers, as indicated by the Expert column in Table 4. The other six answers came directly from the NASA website.

Table 4. NASA Candidate Answers

Q	Candidate Answer	Scaled Support	Righti-ness	Expert
44	12103.6 ± 0.05 km	0.94916	0.97752	Google
44	12103.876 ± 0.80 km	0.56835	0.98037	NASA
2	12755.66 ± 0.80 km	0.64836	1.00000	NASA
14	6793.041 ± 0.80 km	0.72613	0.90019	NASA
10	71491.89 ± 0.40 km	1.00000	0.54538	NASA
36	60267.52 ± 0.40 km	0.67264	1.00000	NASA
36	59867.598 ± 80.47 km	0.21053	0.70557	Google
26	24763.781 ± 2.01 km	0.52801	1.00000	NASA

Ideally, the Scaled Support and Rightiness columns would be highly correlated. This is a very small sample and does not perform very well. In all cases except for the candidate answer to question #10, our Scaled Support value was less than the Rightiness measure. For question #10, we would have chosen this as the best consensus answer, yet it was quite far from the target answer (from Wikipedia).

4.2.2 *Websites with Five or More Answers*

Websites with fewer than five overall responses are omitted from the analysis in Table 5 because the goal is to attempt to identify the relative accuracy of websites across a broad domain of questions. In section 4.5 we analyze the impact of altering that threshold.

Table 5. Top Four Astronomy Candidate Answers

Website Domain	Average Scaled Support	Average Rightiness
spider.seds.org	0.8471	0.6532
www.eightplanetsfacts.com	0.7917	0.6607
creationwiki.org	0.6667	0.6655
nssdc.gsfc.nasa.gov	0.6629	0.8886

Based on all 27 websites that provided candidate answers to at least five of the ten questions, the correlation between Average Scaled Support and Average Rightiness was about 0.61.

The two websites with the highest support (spider.seds.org and www.eightplanetsfacts.com) were not as accurate as other websites. But there is still a strong correlation between Average Scaled Support (predicted accuracy) and Average Rightiness (measured accuracy).

4.3 Overall Relative Reliability

During the development phase, half of the questions were used while the other half were completely untouched. The halves were randomly chosen in advance.

Most of the charts in this section are scatter plots of Average Scaled Support vs. Average Rightiness. In an ideal situation, there would be a line from (0,0) to (1,1) meaning that our calculated Average Scaled Support precisely matches the observed Average Rightiness. As shown, there is a high degree of correlation between the two, which indicates that our Average Scaled Support can be used to predict Average Rightiness.

Overall results (with a 0.81 correlation coefficient) are shown in Figure 2.

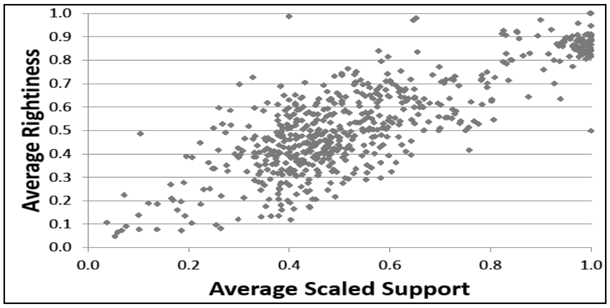

Figure 2. Scaled Support -vs- Rightiness

The apparent cluster in the top right corner is due to the string matching algorithm, described in section 4.4.3.

Visually, there is a strong correlation between our predictive measure (Average Scaled Support) and the observed measure (Average Rightiness). This is reinforced by a high correlation coefficient.

4.4 Relative Reliability by Data Type

There is one chart for each of the data types considered.

4.4.1 *Numeric Results*

Considering the four corners of Figure 3, the bottom left has cases where our predictive measure is low, but also the rightiness is correspondingly low. The bottom right is the worst case, where our confidence is high, but the rightiness is low. The top left is not good either, where we weren't confident, and missed getting some answers. The top right is where we are both relatively confident and relatively correct. There is a large clustering in the center, which is an artifact of averaging both Scaled Support and Rightiness values.

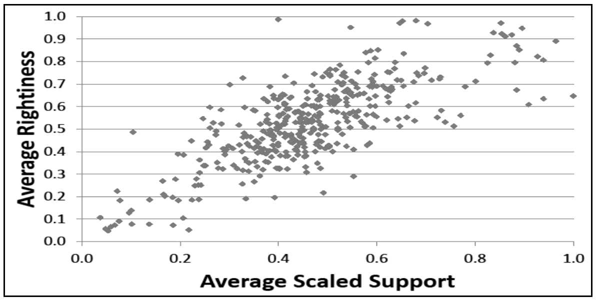

Figure 3. Results for Numeric Questions

The correlation coefficient is about 0.75 for numeric questions.

4.4.2 *Date Results*

The Date matching support algorithm is very precise, which causes the clustering seen in Figure 4. A date like May 6 does not support May 7. Hence, date values tend to be polarized as right or wrong, with not many partial matches. Also, for the domains under consideration (birthdays of President and Baseball players), the concept of implied precision doesn't apply very well. Many websites get these correct, or don't have a candidate answer.

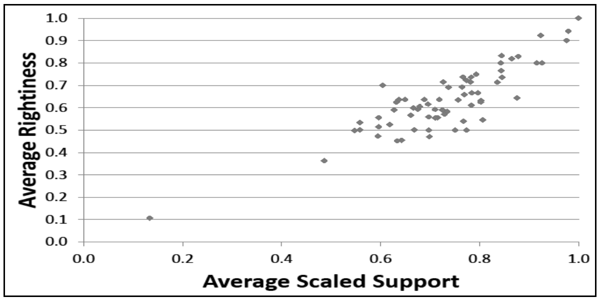

Figure 4. Results for Date Questions

The correlation coefficient is high, about 0.90 for date questions.

4.4.3 *String Results*

The string matching algorithm (Levenshtein's edit distance) tends to have a large number of correct or nearly correct answers, with small edit distances. These are represented by the cluster in the top right corner of Figure 5.

However, they also can get partial overlap between almost any two arbitrary words. Edit distance is only zero if there are no letters at all shared between the two words. Even 'apple' and ‘orange’ share the letters ‘a’ and ‘e’ yielding a net support value of 1/6. These are represented by the large cluster in the middle bottom of the chart. Our Support calculation yields an optimistic value compared to the actual rightiness.

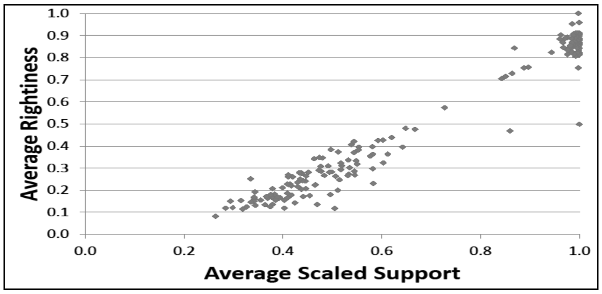

Figure 5. Results for String Questions

The correlation coefficient is very high, about 0.98 for string questions.

4.5 How Many Questions?

Most of the analysis in this paper is based on the requirement that websites provide candidate answers to at least ten questions. In this section, we consider the outcome of varying that number.

The X axis in Figure 6 shows the minimum number of responses, up to 50. The left axis is a $\log_{10}$ scale of the number of websites with candidate answers to at least that many questions. It starts at over 10,000 websites with at least one candidate answer, and descends to just over 100 websites with candidate answers to at least 50 questions.

The correlation coefficient is measured for each of these 50 cases, and ranges from a little more than 0.6 to about 0.9.

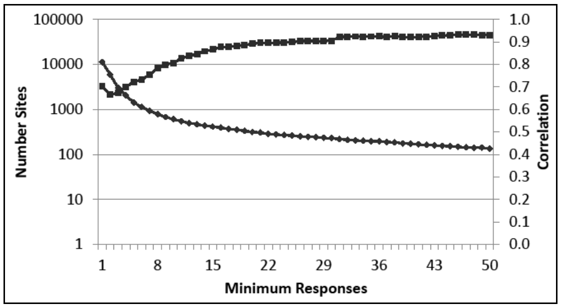

Figure 6. Results by Number of Questions

We conclude that websites with consensus answers to many questions are more likely to have answers closer to the target answer.

5. CONCLUSIONS

We have demonstrated that websites with consensus answers (that is, websites that tend to agree with other websites) are more likely to be reliable. To show this, we analyzed the answers to a corpus of factoid questions from a wide range of websites (using common search engines). The net result is that we can approximately measure the likelihood that a website will have correct answers by comparing its answers to other websites.

6. FUTURE WORK

The three core data types for this analysis can easily be expanded. Numeric subtypes would distinguish real-world measurements (the length of a river) from computed values (e.g., the number of millimeters in a mile). Dates can be expanded to include time-of-day, as well as pre-historic dates. See [9] for additional examples.

There is an assumption in this research that questions have a single answer. This ignores the possibility of multiple meanings for the same question. For example, the diameter of Mars varies depending on equatorial vs. polar measurements. Also, many values are a function of time, such as the population of Paris. See [10] for a framework designed to handle multiple correct answers.

7. REFERENCES

[1] X. Yin, W. Tan and C. Liu, "FACTO: A Fact Lookup Engine Based on Web Tables," in *World Wide Web Conference (WWW)*, Hyderabad, India, 2011.

[2] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems 30,* pp. 107-117, 1988.

[3] X. Yin, J. Han and P. S. Yu, "Truth Discovery with Multiple Conflicting Information Providers on the Web," *Knowledge Discovery and Data Mining (KDD),* 2007.

[4] X. L. Dong, L. Berti-Equille and D. Srivastava, "Integrating Conflicting Data: The Role of Source Dependence," *Very Large Databases (VLDB),* 2009.

[5] A. Galland, A. Marian, S. Abiteboul and P. Senellart, "Corroborating Information from Disagreeing Views," *Web Search and Data Mining (WSDM),* 2010.

[6] S. O'Hara and T. Bylander, "Numeric Query Answering on the Web," *International Journal on Semantic Web and Information Systems,* pp. 1-17, January-March 2011.

[7] B. Katz, S. Felshin, D. Yuret, A. Ibrahim, J. Lin, G. Marton, A. J. McFarland and B. Temelkuran, "Omnibase: Uniform Access to Heterogeneous Data for Question Answering," in *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems*, Stockholm, Sweden, 2002.

[8] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Cybernetics and Control Theory,* pp. 845-848, 1965.

[9] X. Li and D. Roth, "Learning Question Classifiers: The Role of Semantic Information," in *International Conference on Computational Linguistics*, Taipei, 2002.

[10] J. Ko, L. Si and E. Nyberg, "A Probabilistic Graphical Model for Joint Answer Ranking in Question Answering," in *Proceedings of SIGIR*, Amsterdam, 2007.

[11] C. Kwok, O. Etzioni and D. S. Weld, "Scaling question answering to the web," *ACM Transactions on Information Systems,* vol. 19, no. 3, pp. 242-262, July 2001.

[12] M. Barcala, J. Vilares, M. A. Alonso, J. Grana and m. Vilares, "Tokenization and Proper Noun Recognition for Information Retrieval," Departamento de Computacion, Universidade da Coruna, La Coruna, Spain.

[13] D. Roussinov, W. Fan and J. Robles-Flores, "Beyond keywords: Automated question answering on the web," *Communications of the ACM,* vol. 51, no. 9, pp. 60-65, September 2008.