# Numeric Query Answering on the Web

**Steven O'Hara**
*Department of Computer Science*
*University of Texas at San Antonio*
*sohara@cs.utsa.edu*

**Tom Bylander**
*Department of Computer Science*
*University of Texas at San Antonio*
*bylander@cs.utsa.edu*

## ABSTRACT

Query answering usually assumes that the Asker is looking for a single correct answer to the question. When retrieving a textual answer this is often the case, but when searching for numeric answers, there are additional considerations. In particular, numbers often have units associated with them, and the Asker may not care whether the raw answer is in feet or meters. Also, numbers usually denote a precision. In a few cases, the precision may be explicit, but normally, there is an implied precision associated with every number. Finally, we can associate different reliability levels to different sources. We experimentally show that, in the context of conflicting answers from multiple sources, numeric query accuracy can be improved by taking advantage of units, precision, and the reliability of sources.

***Keywords:*** Semantic Web, Text Retrieval, Intelligent Agents, Question / Answering.

## MOTIVATION

Suppose you wanted to know the answer to a specific question, such as "What is the atomic mass of Copper?" or "What is the diameter of Pluto?" Certainly, you could Google (www.google.com) it and get many different answers. You could also ask experts, or go to specific websites that you expect to know an answer, or run software to get an answer. Now, suppose you were able to ask all of these sources at the same time, and wanted to consolidate those answers into a single answer. You'd probably find some answers that are incorrect; and some answers that are more precise than others; and there may even be multiple "correct" answers.

Expecting any single source to have the best answers for a wide range of questions is quite unreasonable. Sources like Wikipedia (www.wikipedia.org), Google Calculator, Yahoo! Answers (answers.yahoo.com), Wolfram Alpha (www.wolframalpha.com) and others are viable sources, but not to the exclusion of the other possible sources. One of the major tenets of this research is that having a range of sources for answers is likely to increase overall accuracy.

Many questions follow this same form: given an object, what is the value of a specified attribute? What is the radius of Mars? What is the population of Rome? What is the length of the Nile River? How tall is Mt. Everest? And so forth. The literature uses the term "factoid" for these types of questions (Lin & Katz, 2003).

With the immense amount of information available on the internet, and the advent of systems like GROK (described herein), it is possible to send a single question and receive many different answers quickly. The GROK architecture provides the mechanism for broadcasting questions to multiple agents (Singh & Huhns, 2005), which in turn search websites for answers. Next, those answers are consolidated into a "best" answer and returned back to the user.

Natural Language processing (such as understanding, paraphrasing or generation) is beyond the scope of this project. See the Text Retrieval Conference website (trec.nist.gov) for some recent research.

## HYPOTHESIS

Numeric query answering is different than ordinary query answering (Katz, et al., 2002), (Kwok, Etzioni, & Weld, 2001), (Lin & Katz, 2003), (Roussinov, Fan, & Robles-Flores, 2008) in two fundamental ways: unit conversions and implied precision. Query accuracy can be improved by taking advantage of these distinct differences. It can also take advantage of source reliability, which is considered in this paper.

### Numeric Values Often Have Units

When searching for numeric values, many of them will be expressed in units, such as feet or meters. When a candidate value is found, the units can be converted automatically. So, if you are searching for the radius of Mars, some answers may come back in meters, some in miles, and some in kilometers. If the search is expecting only miles, it may miss some good answers that were expressed in kilometers, for example.

### Numbers Have Implicit Precision

In a few cases, numbers have explicit precision. Wikipedia, for example, often shows values in the form "2100 ± 40 grams" which means 2060 to 2140 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. These interval values can be compared against each other to see if they reinforce each other.

Naturally, this assumption can fail in both directions. One might find a value such as "1000 meters" that really means exactly 1000.000 meters. And one might find a value such as "123.45 miles" that is not really accurate to one hundredth of a mile. Part of this research is testing to see if overall accuracy improves with this assumption, even knowing that it will occasionally fail.

### Sources Have Different Reliability

Sources like NASA are tightly controlled and very likely to have accurate answers to most astronomy-related questions. Wikipedia is loosely controlled (for a discussion on reliability, see en.wikipedia.org/wiki/Reliability_of_Wikipedia); information is often entered by non-experts and is often invalid. Google, in general, is completely uncontrolled. When comparing answers, it seems reasonable to consider the reliability of the source.

## RELATED WORK

The annual Text Retrieval Conference (TREC) has been the mainstay of question answering systems for many years. Many serious academic efforts have participated, although the majority of

the work assumes a finite corpus of documents that can be pre-processed, unlike the web which is virtually unbounded.

Note that, to the knowledge of the authors, none of the efforts consider the implicit precision associated with numbers, and hence none of them have a concept like Interval Support at described in the Statistical Foundation section, later in this paper.

### Academic Efforts

Numeric query answering has been identified as a different type of problem for some time. For example, (Ferret, Grau, Hurault-Plantet, Illouz, & Jacquemin, 2001) categorizes answer types as Named Entities, such as person, organization, location (city or place), number (a time expression or a number expression).

The earliest web-based question answering system is START (Katz, 1997) which has been in continuous operation since 1993. It has its own knowledge base and is able to search online for answers in sources like the CIA World Fact Book.

The KnowItAll (Cafarela, Downey, Soderland, & Etzioni, 2005) system uses a variety of natural language techniques to collect triples with very high accuracy, although it does require some human guidance. TextRunner (Yates, Banko, Broadhead, Cafarella, Etzioni, & Soderland, 2007) crawls the web looking for object-relationship-object triples that it collects for answering queries. It has a big focus on complete automation – no human input needed to tune rules. Both of these systems share the problem of how to generate the right query for Google, and how to identify extract the candidate answer.

(Clarke, Cormack, & Lynam, 2001) uses an iterative algorithm to validate source reliability, based on candidate answer redundancy and term frequencies. (White, Cardie, & Ng, 2002) creates summaries of multiple websites to highlight discrepancies between them; their focus is on tracking events in which the known facts can change on a daily basis (e.g., the damage and loss of life from an earthquake). TruthFinder (Yin, Han, & Yu, 2007) is based on the idea that a web site is reliable if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many reliable web sites.

FACTO (Yin, Tan, & Liu, 2011) is intended as part of Microsoft's Bing search engine. It also crawls the web looking for object-attribute-value triples and compiles them into a database for retrieval when somebody does a search. Their system utilizes TextRunner to extract facts from the web pages. Their assumption that there exists a single correct answer to put into the database is similar to our goal of identifying the best answer.

### Production Systems

When doing a Google search, sometimes Google will produce a quick answer. Although not published, it appears to be extracting the more frequently occurring answer found in snippets from selected websites. Systems like ask.com and Wolfram Alpha are designed to produce answers to questions, not just links to potentially relevant web sites. Systems like Yahoo! Answers only get answers from other people.

## ARCHITECTURE

The foundation for this research is a system called GROK, which stands for Global Repository of Ordinary Knowledge. GROK was developed by the authors to provide a robust question-answering platform between Askers and Answerers.

In the simplest case, the GROK Client starts with a question, and each of GROK Experts attempts to provide one or more candidate answers. The Experts are typically wrappers for websites such as NASA, Wikipedia or Google.

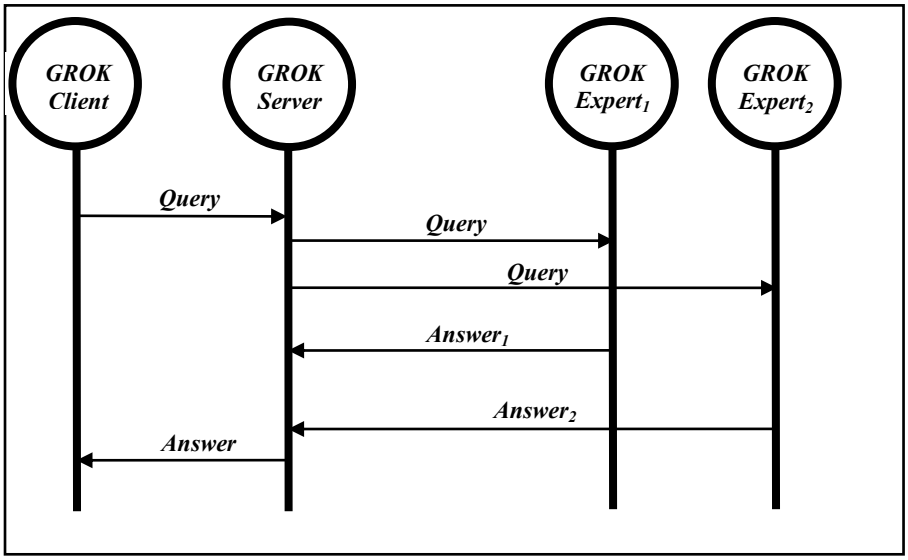


*Figure 1. Simplified GROK sequence diagram*

The preceding figure is simplified, as it omits the shared ontology. It also doesn't really show that there are many Clients, and that the Experts can be agents, software, or humans contacted via email.

### GROK Experts

GROK Experts are registered with the GROK Server for each domain that they have expertise in. Questions will come in from users. Sources like Google and Wikipedia will have broad knowledge covering all domains. Some questions might be answerable through software run on the expert's computer, while others might be answered by an email sent to a human expert.

### GROK Google Expert

The GROK experts that we used for this research are tuned to extract information in an expected form, from a specific website. GROK can also ask human experts and consult the output of running software, but that is not used for our results in this paper. For example, the NASA website has tables full of answers for astronomy questions. In some cases, such as Wikipedia, they have Resource Description Framework descriptors to access the information directly, but in most cases, a custom wrapper has to be created to extract the information from the web page directly.

The GROK Google expert is different. It generates a series of queries, as if the user typed them, and examines the two or three line snippets returned for each match. All questions are processed by this expert.

## GROK Expert Confidence Levels

This analysis considers four different levels of reliability, or confidence.

*Table 1. Expert Confidence Levels*

| Level | Scale Factor | Example |
|---|---|---|
| High | 300 | NASA |
| Medium | 100 | Wikipedia |
| Low | 30 | Google – full pattern |
| Very Low | 0.5 | Google – no pattern |

The scale factor only has meaning relative to the other scale factors. For example, a single candidate answer with High level of confidence (300) is equivalent to 10 candidate answers with Low confidence (30 each). These ratios are arbitrary; additional research is needed to tune these ratios. Clearly, websites are not equally likely to have correct answers. There is no a priori way to exclude websites, nor are there any websites that are guaranteed to be correct.

Google responses can only have two confidence values. When a match is made for the full pattern, such as "The diameter of Pluto is …", the confidence is considered to be Low. If less than 10 candidate answers are found this way, an additional query is attempted with just the object and attribute, such as "diameter Pluto". Any candidate answers returned for this query are considered to be Very Low confidence, because it looks through the entire snippet for any numbers followed by a valid unit. A candidate answer might not be referring to Pluto or its diameter, but the answer will still have a distance unit in it.

A research project has been initiated by the authors to quantify website confidence. The idea is to create a large collection of Questions across different domains. Send all questions through the GROK system collect all candidate answers. Reduce the web site sources to their DNS name (www.nasa.gov for example) instead of the full URL. Now, collect all the candidate responses for each DNS name and assess an average correctness for them based on the methodology from this research paper. This measure becomes an indicator for website confidence. Preliminary results are showing that websites like www.nasa.gov are showing very high confidence, confirming expectations.

## GROK Server

The GROK server knows what information is available, but doesn't actually know anything. The best way to think of this is that the server knows the column names and types of all the tables in the database, but none of the actual data in the tables. It knows about domains, and which experts can answer questions in each domain. It is really just a broker that distributes questions and collects candidate answers. These candidate answers are then consolidated into a single consensus answer that is delivered back to the GROK Client.

## GROK Clients

Users may wish to find answers to domain-specific questions. Perhaps computer programs need information as well, such as weather forecasts. The GROK system can be used to distribute the question to multiple Experts. GROK will consolidate and normalize answers from multiple experts, into a single consensus answer.

In this research, the GROK Client is a program that reads a prepared list of questions spanning several different domains.

## Shared Ontology

When questions are posed to the GROK system, they are set in the context of a shared ontology. The idea is similar to Cyc (Matuszek, et al., 2005), but this ontology is much less detailed. When a user wishes to know the diameter of Pluto, for example, the following ontology snippet is shared between the GROK components:

```
  <type name="CelestialObject">
    <attribute name="Name"
        type="String" />
    <attribute name="Radius"
        type="Distance" />
    <attribute name="Diameter"
        type="Distance" />
    <attribute name="Mass"
        type="Mass" />
  </type>
  <type name="Planet"
        parent="CelestialObject">
    <attribute name="WhichStar"
        type="Star" />
    <attribute name="HasRings"
        type="Boolean" />
  </type>
```

*Figure 2. Ontology for Planet*

Topics are very high level concepts, and follow the object-oriented pattern of inheritance. Objects and attributes have to be from the types and attributes in the ontology that is shared between all users and experts. Note that there are no instances of objects in the ontology. E.g., Pluto is not a known concept.

## Questions and Answers

Questions are represented using plain text XML. Here is a sample question.

```
<question qtopic="Astronomy"
    qclass="Planet"
    qobject="Pluto"
    qattribute="Diameter" />
```

*Figure 3. Sample XML question*

Each expert receives questions like this, and replies with a list of XML answers, for example:

```
<answer avalue="2337.5"
    adelta="37.5"
    aunits="km"
    aurl="http://www.pd.astro.it/
      E-MOSTRA/NEW/A2024PLU.HTM" />
```

*Figure 4. Sample XML answer*

A variety of methods are available to communicate these XML messages between the GROK components. For this research, simple TCP/IP sockets are used.

### Units Conversion

Another part of GROK is the automatic conversion of units, e.g., from miles to kilometers, which is processed inside the GROK server. The following snippet shows how the Distance type is represented.

```
<measures>
  <type name="Distance" base="foot">
    <unit name="mile" abbrev="mi"
        factor="5280"/>
    <unit name="meter" abbrev="m"
        factor="3.280839895"/>
    <unit name="kilometer" abbrev="km"
        factor="3280.839895"/>
    <unit name="kilometre"
        factor="3280.839895"/>
    <unit name="foot" plural="feet"
        factor="1"/>
    <unit name="inch" plural="inches"
        factor="0.08333333333"/>
```

*Figure 5. Sample unit conversions*

In the future, this unit conversion processing could be handled through some form of internet web service.

## STATISTICAL FOUNDATION

Suppose one had two numbers in hand, like 210 and 211. We assume that the first value has an implied range of 205 to 215, while the second has an implied range of 210.5 to 211.5. In order to determine what answer to use as the consensus answer, it is necessary to compare them in some fashion. We describe our technique and show that it is identical to a probability model.

### Interval Supports

In our system, comparing two numbers is based on converting them into intervals and computing the support (overlap) between them. Given any two intervals along a continuous line, there are several different ways one can support the other. Call these intervals A and B. A ranges from $A_{LO}$ to $A_{HI}$, and B from $B_{LO}$ to $B_{HI}$. All intervals are non-trivial, i.e., $A_{HI} > A_{LO}$ and $B_{HI} > B_{LO}$. There are four possibilities:

(1) A is inside B. ($B_{LO} \leq A_{LO} < A_{HI} \leq B_{HI}$).

(2) B is inside A. ($A_{LO} \leq B_{LO} < B_{HI} \leq A_{HI}$).

(3) A is disjoint from B. ($A_{HI} < B_{LO}$ or $A_{LO} > B_{HI}$).

(4) A partially overlaps B. ($A_{LO} \le B_{LO} \le A_{HI} \le B_{HI}$ or $B_{LO} \le A_{LO} \le B_{HI} \le A_{HI}$).

If A is totally inside B, then it supports B 100%, which is equivalent to asserting that A implies B. If A and B are disjoint, then the support is zero. If there is partial overlap between A and B, then there is partial support. Here is a definition of the 'supports' relationship:

A supports B = 0 if $A_{HI} < B_{LO}$ or $A_{LO} > B_{HI}$.

Otherwise, A supports B = $$\frac{\min(A_{HI}, B_{HI}) - \max(A_{LO}, B_{LO})}{A_{HI} - A_{LO}}$$

Consider the following chart, looking at the diameter of Pluto, in miles. The horizontal axis represents the candidate answers given in response to a Google query, one answer per hash mark.

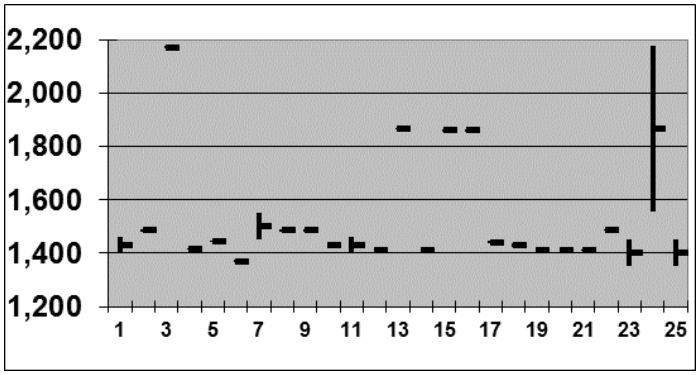


*Figure 6. Diameter of Pluto responses*

Here is the same chart, zoomed in on the vertical axis to 1350 to 1500 miles.

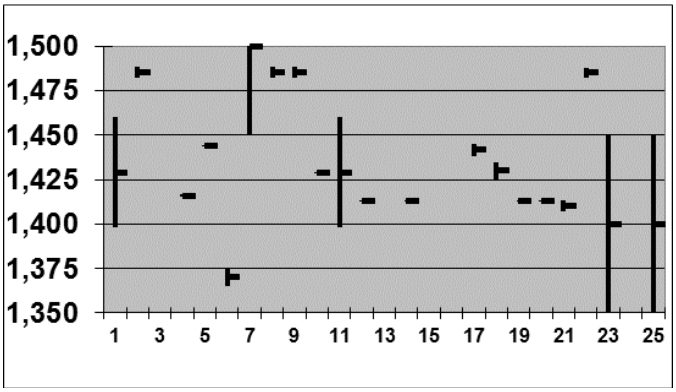


*Figure 7. Diameter of Pluto, zoomed view*

Each small horizontal bar shows a candidate answer, and the vertical bar shows the implied interval for that answer. For example, the first candidate answer was given as 2300 km, which implies 2250 to 2350 km, which is the same as 1429±31 miles. Which is 1398 to 1460 miles, as labeled "1" in the chart. Here are the 25 candidate answers for these charts:

*Table 2. Pluto Diameter Candidate Responses*

| # | Value | Units | Miles | Delta |
|---|---|---|---|---|
| 1 | 2300 | KM | 1,429 | 31.1 |
| 2 | 2390 | KM | 1,485 | 3.1 |
| 3 | 2170 | MILES | 2,170 | 5.0 |

| 4 | 1416 | MILES | 1,416 | 0.5 |
|---|---|---|---|---|
| 5 | 1444 | MILES | 1,444 | 0.5 |
| 6 | 1370 | MILES | 1,370 | 5.0 |
| 7 | 1500 | MILES | 1,500 | 50.0 |
| 8 | 2390 | KM | 1,485 | 3.1 |
| 9 | 2390 | KM | 1,485 | 3.1 |
| 10 | 1429 | MILES | 1,429 | 0.5 |
| 11 | 2300 | KM | 1,429 | 31.1 |
| 12 | 2274 | KM | 1,413 | 0.3 |
| 13 | 1865 | MILES | 1,865 | 2.5 |

| # | Value | Units | Miles | Delta |
|---|---|---|---|---|
| 14 | 2274 | KILOMETERS | 1,413 | 0.3 |

| 15 | 1860 | MILES | 1,860 | 5.0 |
|---|---|---|---|---|
| 16 | 1860 | MILES | 1,860 | 5.0 |
| 17 | 2320 | KM | 1,442 | 3.1 |
| 18 | 1430 | MILES | 1,430 | 5.0 |
| 19 | 2274 | KILOMETERS | 1,413 | 0.3 |
| 20 | 2274 | KILOMETERS | 1,413 | 0.3 |
| 21 | 2270 | KM | 1,411 | 3.1 |
| 22 | 1485 | MILES | 1,485 | 2.5 |
| 23 | 1400 | MILES | 1,400 | 50.0 |
| 24 | 3000000 | METERS | 1,864 | 310.7 |
| 25 | 1400 | MILES | 1,400 | 50.0 |

When two answers, including their implied ranges, have a partial overlap, the amount of support is computed to a value between zero and one. For example, consider the first candidate answer (call it A) and the amount of support given by the fifth candidate answer (call it B), which is 1500 miles, with an implied range of 50. So

$A_{LO} = 1398$

$B_{LO} = 1450$

$A_{HI} = 1460$

$B_{HI} = 1550$

Support = $(A_{HI} - B_{LO}) / (A_{HI} - A_{LO}) = (1460 - 1450) / (1460 - 1398) = 0.16129$

So the 2300 km answer supports the 1500 mile answer about 16%.

## Conditional Bayes Probabilities

The above 'supports' relation is equivalent to a traditional Bayes conditional probability, Prob (A | B) = Prob (A and B) / Prob (B), if some simplifying assumptions are made. The first assumption is that an interval is equivalent to a uniform distribution. The second is that the range of possible values is continuous. That is, the assertion "the value is in the interval A" is equivalent to asserting a uniform distribution, "the value is equally likely to be any value in the interval A".

To show the equivalence between the Bayes calculations and the Support calculations, we will make one further assumption, namely that the values come from a finite range of possible values. Let us call that range U, with values ranging from $U_{LO}$ to $U_{HI}$. All values must lie within U.

$Prob(B) = (B_{HI} - B_{LO}) / (U_{HI} - U_{LO})$

$Prob(A \text{ and } B) = Overlap / (U_{HI} - U_{LO})$

The amount of overlap is the smaller of the high values minus the larger of the low values. So,

$Overlap = \min(A_{HI}, B_{HI}) - \max(A_{LO}, B_{LO})$

Hence, $Prob(A \mid B) = [\min(A_{HI}, B_{HI}) - \max(A_{LO}, B_{LO})] / (A_{HI} - A_{LO})$

Which is exactly as shown in the Interval Supports section above.

## Ranking 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 *receive* from the other candidate answers, rather than the amount of support they *give*.

The ranking of the candidate answers is based on computing the support between all pairs of answers, and adding them up for each answer. The candidate answer that gives the most support to the other answers is chosen as the "best" or consensus answer. Note that the opposite approach is valid as well, choosing the candidate answer that receives the most support. The difference is that the former approach tends to find a more precise answer while the latter tends to find very broad answers.

## EXAMPLE ANALYSIS

To illustrate the analysis, the following two questions will be studied in some detail. The Q column is just a reference question number.

*Table 3. Sample questions*

| Q | Object | Attribute | Target Answer (*) |
|---|---|---|---|
| 1 | Aland Islands | Total Area | 13,517 sq km |
| 2 | Amazon River | Length | 6,992 km |

(*) These target answers were extracted from the 2008/9 Wikipedia Selection for Schools.

Below are the candidate answers, all from Google. Note that, in the ontology, atomic radius of an element is defined to have an absolute possible upper limit of 1,000 picometers. This applies to all elements and is just a mechanism to reduce bizarre candidate answers. For question 26 below, seven candidate answers were eliminated because they far exceeded this limit.

Here are the 23 candidate answers for question 1 (What is the Total Area of the Aland Islands?).

*Table 4. Responses for question 1*

| | |
|---|---|
| 1512 square kilometers (*) | 1512 square kilometers (*) |
| 584 sq mi (+) | 583 sq. mi (+) |
| 970 square miles | 1512 square kilometres (*) |
| 1512 square kilometres (*) | 583 sq. mi (+) |
| 583 sq. mi (+) | 1512 square kilometres (*) |
| 1512 square kilometers (*) | 583 sq. mi (+) |
| 583 sq. mi (+) | 1512 square kilometres (*) |
| 1512 square kilometers (*) | 583 sq. mi (+) |
| 584 sq mi (+) | 1512 square kilometres (*) |
| 1426 sq. km. | 583 sq. mi (+) |
| 1426 sq. km. | 1512 square kilometers (*) |
| 1426 sq. km. | |

Based on the preceding analysis, the consensus answer of 1512 sq. km was chosen because it has the most support from other candidate answers. In addition to the ten answers (marked with *) of

1512 sq. km., the 583 and 584 sq. miles answers (marked with +) are all partially supported by the answer of 1512 sq. km.

Question 2 (What is the Length of the Amazon River?) got 40 candidate answers:

*Table 5. Responses for question 2*

| | | |
|---|---|---|
| 3 m | 0 in | 3969 mi (+) |
| 1600 km | 4050 mi | 5597 m |
| 1000 mi | 1609 m | 0 m |
| 4380 km | 6387 km (*) | 2300 miles |
| 2720 mi | 3969 mi (+) | 6450 km |
| 110 feet | 5597 m | 4000 mi (+) |
| 4380 km | 4380 km | 6387 km (*) |
| 2720 mi | 2720 mi | 3969 mi (+) |
| 4170 miles | 6516 kilometers | 5597 m |
| 6296 km | 6400 km (+) | 6762 km |
| 8 feet | 3977 mi | 4202 mi |
| 3 m | 7050000 km | 5597 m |
| 3 m | 2722020 mi | |
| 3 m | 6387 km (*) | |

The candidate answer (marked with *) of 6387 km supports the most answers. It is only given in three of the answers, but it supports five other answers (marked with +).

### Comparing Answers

The consensus answer is first converted to the same units as the target answer. To compare a consensus answer to the target answer in our evaluation, we define a distance measure. Let the consensus answer be $A_{LO}$ to $A_{HI}$, with $A_{MID}$ halfway between. Similarly, let the target answer be $B_{LO}$ to $B_{HI}$, with $B_{MID}$ halfway. Then, ansDistance is defined as:

$$\frac{|A_{LO} - B_{LO}| + |A_{MID} - B_{MID}| + |A_{HI} - B_{HI}|}{3|B_{MID}|}$$

ansDistance is a simplified version of the average distance between corresponding points on the intervals. Consider an example, where the consensus answer (A) is 80±40 and the target answer (B) is 50±30.

ansDistance = average(|40-20|, |80-50|, |120-80|) / 150 = 0.20

## BASELINE AND THREE EXPERIMENTAL CONDITIONS

The baseline experimental condition includes unit conversion, implicit precision, and different confidence levels for different sources. The baseline is expected to have the best performance.

Condition 1 (no Unit Conversions, "No Conv.") does not allow candidate answers in different units to support each other. For example, answers in miles and kilometers are valid, but are prevented from supporting each other.

In condition 2 (no Implied Precision, "No Impl."), all values are considered to be totally precise. For example, 4001 miles does not support 4000 miles under this condition.

Condition 3 (no Confidence Levels, "No Conf.") does not consider the confidence of the source. E.g., an answer from Google is given the same weight as an answer from NASA. For the Wikipedia Development Phase, the only source activated was Google, so the only confidence levels were Low and Very Low. (Low corresponds to a Google search for a particular way of phrasing the answer, while Very Low corresponds to a keyword search.)

The core of the research starts with a baseline configuration consisting of all available logic. Our hypothesis is that the baseline should have the best overall performance. Then three experimental conditions are tried, where each hypothesis is disabled one at a time.

*Table 6. Baseline and experimental conditions*

| Experimental Conditions: | Unit Conversions | Implied Precision | Confidence Levels |
|---|---|---|---|
| Baseline | ✓ | ✓ | ✓ |
| 1. No Conv. | × | ✓ | ✓ |
| 2. No Impl. | ✓ | × | ✓ |
| 3. No Conf. | ✓ | ✓ | × |

There were three phases to this research. The first phase was used to get the functionality operational and computations validated. The second phase was a detailed study into the three conditions described below. The third phase was the testing phase and was used to validate the results of the second phase.

### Phase 1: Initial Development

A total of 115 questions were used for the initial development phase. Here are some representative questions:

- "What is the diameter of Pluto?"
- "What is the radius of Pluto?"
- "How many grams are in a ton?"
- "How far is it to Miami from Chicago?"
- "What is the atomic mass of Copper?"
- "What is the length of the Nile?"
- "What is the area of Vietnam?"
- "What is the area of Alabama?"

During the initial development phase, considerable emphasis was placed on the approach. I.e., making sure all the statistics were calculated correctly, all the answers were consolidating correctly into a single consensus answer, etc. Also, the wrappers for sources like Google, Wikipedia, the CIA World Fact Book (www.cia.gov/library/publications/the-world-factbook), etc. were all implemented and tested.

### Phase 2: Wikipedia Development

Wikipedia has developed a secondary school online encyclopedia. It can be downloaded and used for research purposes. This website was used to create a list of 1,418 numeric questions and target answers. Of the 1,418 questions, 278 did not return enough answers for analysis, leaving a total of 1,140 questions.

The following chart indicates the performance of each condition. The y axis is the count of the questions in that accuracy range. The accuracy values are based on the ansDistance distance measure, so values closer to zero imply greater accuracy. A value of zero means the experimental condition got precisely the target Wikipedia answer. Values in the range 0 to 1/4 are within 25% of the target answer, and a value of "One or more" means it was off by 100% or more.

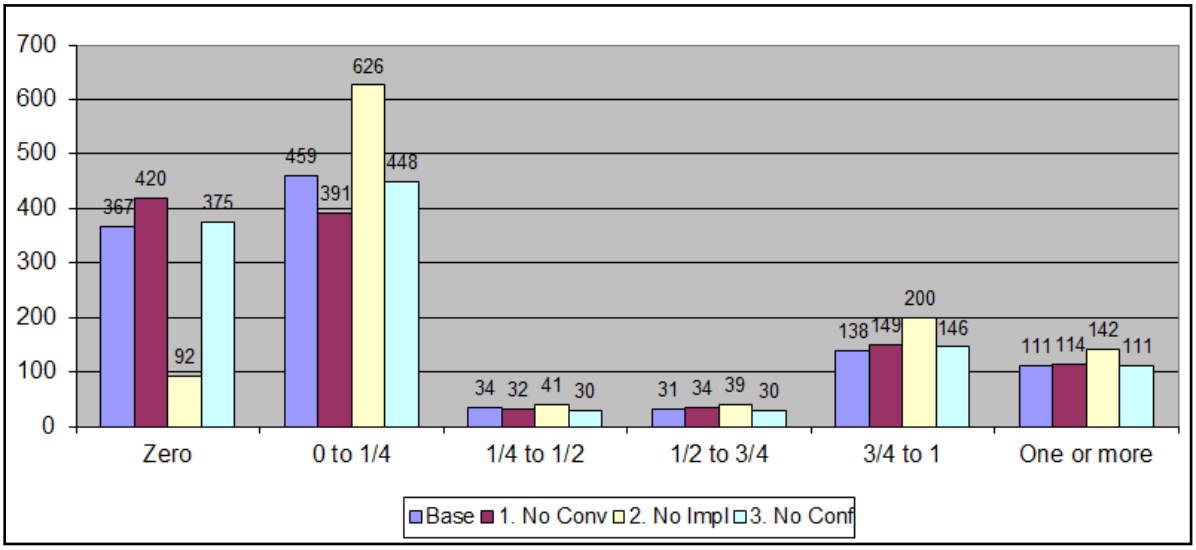


*Figure 8. Wikipedia Development Phase accuracy*

For example, Condition 3 (no confidence) got 448 of the 1,140 questions slightly wrong, but still within 25% of the target answer. Overall, the baseline and each of the conditions got a large fraction of the answers within 25% of the target answer. The rightmost two groups of bars (3/4 to 1 and One or more) respectively represent cases where the consensus answer is far from the target. Usually, an answer in the "3/4 to 1" group corresponds to where the consensus answer is very small relative to the target answer (can't be off by more than 100%). The "One or more" group corresponds to where the consensus answer is very large relative to the target answers (more than twice as much).

It is interesting to note that Condition 1 (No Conv, no conversion of units) actually got more answers exactly right (420 to 367). Possibly this is because it introduces no errors from unit conversion. The total of the exact plus close (0 to 1/4 including Zero) is 826 for baseline and 811 for condition 1, which indicates that the baseline condition was comparable to condition 1.

It is also interesting to note that condition 2 (No Impl, no implied precision) did not get very many answers exactly the same as the target answer (92), but got quite a few answers close to it (626).

In order to compare each condition to the baseline result, a Paired Difference test is performed. The difference between the Baseline ansDistance and the experimental condition ansDistance is computed. Roughly ansDistance corresponds to (consensus-target)/target, which means that a negative difference between the baseline ansDistance and a condition ansDistance means that the baseline predicted a better answer. However, since values can be wildly different, even off by orders of magnitude in some cases, the difference is limited to the range -1 to 1.

Based on all 1,140 questions, the Paired Difference computation follows:

*Table 7. Paired difference results, Wikipedia Phase*

| Computation | Cond 1. No Conv. | Cond 2. No Impl. | Cond 3. No Conf. |
|---|---|---|---|
| Avg. Diff | -0.0109 | -0.0854 | -0.0042 |
| Sample Var. | 0.0549 | 0.1722 | 0.0104 |
| T Statistic | -6.70 | -16.74 | -13.71 |

These T test values are very significant and show that the average difference was negative with very high confidence ($p << .001$). That is, the Baseline outperformed the three conditions on this measure.

Likewise, a Signed Z test is performed. It is similar to the Paired Difference Test, but only the sign of the difference is considered (possible values -1, 0 or 1). A value of -1 means the baseline got a better answer than the experimental condition. A value of 0 means both were equally accurate, and 1 means the baseline got a worse answer than the experimental condition.

Based on all 1,140 questions, the signed Z test computation follows:

*Table 8. Signed Z test results, Wikipedia phase*

| Computation | Cond 1. No Conv. | Cond 2. No Impl. | Cond 3. No Conf. |
|---|---|---|---|
| Baseline worse | 192 | 257 | 49 |
| Baseline better | 152 | 691 | 51 |
| Z = signed test | 2.16 | -14.10 | -0.20 |

The Z value for the first condition shows a fairly high certainty that the Condition 1 had better performance than the Baseline ($p < .05$). The second condition is shown to have worse performance than the Baseline ($p << .001$). But the third condition has nearly the same performance as the baseline.

### Phase 3: Testing

Two major sources were used for the testing phase. No code changes were made for this phase; all processing was done using the techniques developed during the previous phase. The questions were taken from very different domains to validate the results.

*Table 9. Test phase question counts*

| Topic | Class | Attribute | Total |
|---|---|---|---|
| Chemistry | Molecule | Temperature | 310 |
| " | " | Weight | 647 |
| Geography | County | Area | 3,143 |
| " | State | Area | 52 |
| Total | | | 4,152 |

The "target answer" source for the chemistry (molecular) data was www.reciprocalnet.org while the source for the geographic information was quickfacts.census.gov. It is difficult to judge the true accuracy of these sources. Of the 4,152 questions, 159 did not return any candidate answers at all in the baseline or one or more of the three conditions, so they were excluded from the anal-

ysis. Leaving a total of 3,993 questions. The average number of candidate answers was about 5.8 and the maximum was 101 (for the Area of Texas).

The following chart indicates the performance of each condition. As before, a value of zero means the experimental condition got precisely the target answer. Values in the range 0 to 1/4 are within 25% of the target answer. A value of "One or more" means it was off by 100% or more.

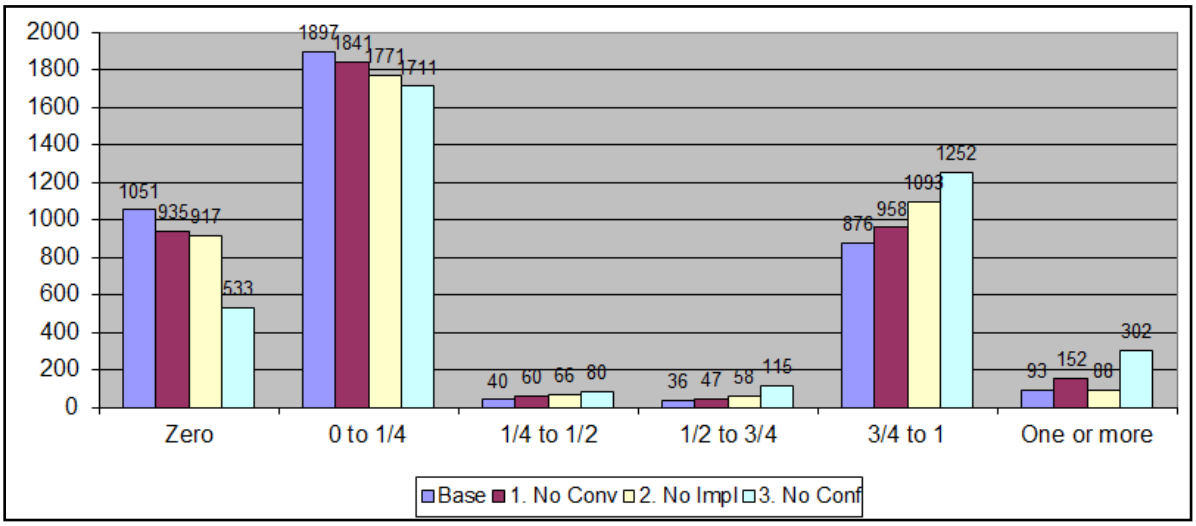


*Figure 9. Test phase accuracy*

The rightmost two groups of bars (3/4 to 1 and One or more) respectively represent cases where the consensus answer is likely very small relative to the target answer (can't be off by more than 100%) or is very large relative to the target answers (more than twice as much).

In this case, the number of answers that approximately match the target answer (0 to 1/4 group) is roughly double the number of answers that exactly match (Zero group). There is also an apparent tendency to get very small answers relative to the target answer (3/4 to 1) compared to answers that are too large (One or more). Both of these observations are a direct consequence of the chosen question set, which is dominated by questions about the areas of counties. Areas of counties are difficult to define precisely due to factors like bordering waters, which helps justify the first observation.

Also, when a county area answer is incorrect, we observed that it is much more likely to be a smaller area than a larger area.

In order to compare each condition to the baseline result, a Paired Difference test is performed. The difference between the Baseline ansDistance and the conditional ansDistance is computed. However, since values can be wildly different, even off by many orders of magnitude in some cases, the difference is limited to the range -1 to 1. A negative difference indicates that the baseline had a more accurate answer.

Based on all 3,993 questions, the Paired Difference computation follows:

*Table 10. Paired difference results, Test phase*

| Computation | Cond 1. No Conv. | Cond 2. No Impl. | Cond 3. No Conf. |
|---|---|---|---|
| Avg. Diff | -0.0385 | -0.0586 | -0.1240 |
| Sample Var. | 0.0738 | 0.1160 | 0.1296 |
| T Statistic | -33.00 | -31.89 | -60.45 |

The Baseline outperformed all the other conditions on the paired difference test ($p << .001$).

Likewise, a Signed Z test is performed. It is similar to the Paired Difference Test, but only the sign of the difference is considered (possible values -1, 0 or 1).

Based on all 3,993 questions, the Signed Z test computation follows:

*Table 11. Signed Z test results, Test phase*

| Computation | Cond 1. No Conv. | Cond 2. No Impl. | Cond 3. No Conf. |
|---|---|---|---|
| Baseline worse | 324 | 1224 | 76 |
| Baseline better | 673 | 2022 | 691 |
| Z = signed test | -11.05 | -14.01 | -22.21 |

The Z value for all three conditions shows a fairly high certainty that each condition had a significant impact on performance. Again, the baseline outperformed each of the experimental conditions ($p << 0.001$).

### Experimental Conclusions

There were mixed results from the Signed Z test from our Wikipedia Development phase, however the other three statistical tests validate our hypotheses, in particular, the Paired Difference Test and Signed Z Test from our Test phase. It is clear that numeric answers should be treated with implied precision. In addition, unit conversion and the reliability of sources are important to take into account.

## KNOWN ISSUES AND FUTURE RESEARCH

### Multiple Values

When talking about real world values, especially those with units attached to them, it is very often the case that the value depends on what is being measured. Consider, for example, the length of a year. One might say 365 days, or 365.25 or 365.2425, or some variation on that. But, in fact, there are many different ways of measuring a year. Wikipedia, for example, lists 11 different kinds of astronomical years (see en.wikipedia.org/wiki/Year#Astronomical_years)! A Julian year has exactly 365.25 days, while a Sidereal year is just slightly longer. Note also, that the length of year varies ever so slightly due to many small factors like solar winds and lunar gravity.

Consider also the diameter of Jupiter. Does that refer to the equatorial diameter? One of the many polar diameters? An average or median diameter? Jupiter may not even have a well-defined surface. “Jupiter is a giant ball of gas and liquid with little, if any, solid surface. Instead, the planet's surface is composed of dense […] clouds”, according to NASA (see www.nasa.gov/worldbook/jupiter_worldbook.html).

### Judging Source Reliability

In this experiment, it was necessary to quantify relative source reliabilities. The ratios between High, Medium, Low, and Very Low are derived empirically, but are arbitrary. Different ratios will influence the results. It is especially difficult to assign reliability to human sources, without some form of feedback system. We intend to incorporate ideas from TextRunner (Yin, Han, & Yu, 2007) into future versions of our system.

### Duplicate Websites

In the process of analyzing multiple sources to see how well they support each other, the results can be easily contaminated by duplicate websites. It seems that many websites copy-and-paste values from other more reputable websites. The act of copying from a source probably indicates that the source is somewhat reliable, but this is not a principled analysis. In this experiment, no attempt was made to reconcile sources to see if they were clones.

Considerable research is already being done is this area. See for example (Galland, Marian, Abiteboul, & Senellart, 2010). (Dong, Berti-Equille, & Srivastava, 2009) is especially interesting in tracking the dependency history of redundant website information over time.

### Dates, Non-numeric Values and Missing Units

Strings and dates can have similar analysis. One might consider Brisbane and Brisbaine to support each other for the capital of Australia. Also, days clearly support months, for example July 4, 2009 clearly supports July 2009. But the quantification of the support is slightly different. Consider also the population of Spain, which usually does not have any explicit units. The analysis in this experiment is applicable, other than Units Analysis.

### Second Place, Third Place etc.

This experiment assumes that a single, best answer is expected. It should be possible to return a ranked list of answers, each with a relative confidence to the others.

### Adding More Knowledge Sources

Sites such as Yahoo! (www.yahoo.com) and Google both have instant answer sources. They have a higher reliability than just a Google search and should be considered as sources. There are also many domain specific sites that can be used for specialized queries. At the moment, adding a new source means adding a new wrapper with custom code. It is possible that patterns will emerge to facilitate the simpler extraction of values from websites. At the moment, writing a new wrapper involves a considerable amount of custom code and testing.

Systems like FACTO (Yin, Tan, & Liu, 2011) scan through web sites looking for attribute-value tables. It is possible that this approach can be used to help extract useful information from websites, reducing the need for manual wrapper creation.

## CONCLUSIONS

The GROK architecture provides a good framework for experimenting with numeric question answering on the web. It allows candidate answers to come from different sources, including Google, Wikipedia, NASA, CIA World Fact Book, and many others. Furthermore, it is able to collect and consolidate multiple answers into a single consensus answer, and then to compare that against a target answer.

The three main hypotheses suggested in the research were all validated. Each of the following was demonstrated to improve performance:

- doing unit conversions,
- using an implied precision for numbers, and
- considering the confidence of the source.

These results were achieved during the Wikipedia Development phase, and validated during the Test phase.

## REFERENCES


Cafarela, M. J., Downey, D., Soderland, S., & Etzioni, O. (2005). KnowItAll: Fast, scalable information extraction from the Web. *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, (pp. 563-570).

Clarke, C., Cormack, G. V., & Lynam, T. R. (2001). Exploiting Redundancy in Question Answering. *Proceedings of the 24th Annual International ACM Conference on Research and Devel-opment in Information Retrieval (SIGIR).*

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

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., & Jacquemin, C. (2001). Terminological Variants for Document Selection and Question/Answer Matching. *Proceedings of the workshop on Open-domain question answering - Volume 12 (ODQA).* Association for Computational Linguistics.

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

Katz, B. (1997). Annotating the World Wide Web Using Natural Language. *Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet.*

Katz, B., Felshin, S., Yuret, D., Ibrahim, A., Lin, J., Marton, G., et al. (2002). Omnibase: Uniform Access to Heterogeneous Data for Question Answering. *Proceedings of the 7th International Workshop on Applications of Natural Language to Information Systems.* Stockholm, Sweden.

Kwok, C., Etzioni, O., & Weld, D. S. (2001). Scaling question answering to the web. *ACM Transactions on Information Systems, 19*(3), 242-262.

Lin, J., & Katz, B. (2003). Question Answering from the Web Using Knowledge Annotation and Knowledge Mining Techniques. *Proceedings of Twelfth International Conference on Information and Knowledge Management.* New Orleans, Louisiana: MIT Computer Science and Artificial Intelligence Laboratory.

Matuszek, C., Witbrock, M., Kahlert, R. C., Cabral, J., Schneider, D., Shah, P., et al. (2005). Searching for Common Sense: Populating Cyc™ from the Web. *Proceedings of the Twentieth National Conference on Artificial Intelligence.* Pittsburgh, Pennsylvania.

Roussinov, D., Fan, W., & Robles-Flores, J. (2008, September). Beyond keywords: Automated question answering on the web. *51*(9), 60-65.

Singh, M., & Huhns, M. (2005). *Service-Oriented Computing: Semantics, Processes, Agents.* West Sussex, England: John Wiley & Sons, Ltd.

White, M., Cardie, C., & Ng, V. (2002). Detecting Discrepancies in Numeric Estimates Using Multidoc-ument Hypertext Summaries. *Proceedings of the Second International Conference on Human Language Technology Research (HLT).*

Yates, A., Banko, M., Broadhead, M., Cafarella, M., Etzioni, O., & Soderland, S. (2007). TextRunner: Open Information Extraction on the Web. *NAACL HLT Demonstration Program* (pp. 25-26). Rochester, NY: Association for Computational Linguistics.

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

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