An improved approach to geographically locating web clients

Howard Kim and Simon Dobson

Department of Computer Science, Trinity College Dublin, Ireland
howard.kim@cs.tcd.ie

Abstract. Many modern web applications need access to information about users' real-world locations, in order to provide services such as tailored advertising. Naive approaches to locating users rarely exceed 40% accuracy due to the widespread use of non-geographically-bound domain names. We describe an approach to location that improves accuracy to around 80%. The approach is completely transparent to users, is scalable and robust, and may be simply deployed as a service within an application container.

Introduction

The Internet is a borderless network of machines interconnected all over the world. It does not adhere to the physical boundaries that exist in the real world where borders state where one country ends and another begins. However, it is becoming increasingly important to geographically identify where a user is located when accessing a given web site, because more and more web sites are running into difficulties about international laws and customs. The term geolocation has been used to describe geographical identification [4].

As an example recently human rights organisations in France protested about Yahoo!'s auction site selling Nazi memorabilia to French citizens. A French court ordered Yahoo to devise a technical solution that would block French Internet users from accessing the Nazi related material posted to the portal's American auction site. Yahoo was fined $13,000 per day while French users could access the site. Similar cases in Italy and Germany have been brought against Internet sites [4].

But laws are not the only problem. The World Wide Web is suffering from it's own success; it has now reached over 375 million users, in over 200 countries [3]. But not everyone speaks the World Wide Web's dominant language "English". For the Internet to become truly global web sites must be able to adapt to users languages and custom preferences.

Every machine on the Internet is assigned a unique address called an Internet Protocol (IP) address. The main problem about IP addresses in relation to geolocation is that IP addresses are location transparent, this means that an IP address can be located in any country; usually IP addresses are allocated in blocks. There may be situations where a company in America acquires a block of addresses but uses these addresses for machines in Ireland. The difficulty in identifying users is increased with there being little or no support for geographical identification currently built into the Internet's architecture.

Geolocation technology is only in nascent stages yet businesses can use geolocation in many ways:

1. Localised content

Internationalised web sites can be localised to given languages and customs. Also content can be blocked or restricted when laws differ in countries; this has great appeal to companies operating in multiple areas. A geolocation service can stretch the boundaries or lines in this virtual world to their needs.

2. In e-commerce

Web sites think of users globally yet act on a local level. Visitors to such sites enjoy a more meaningful and personalised experience. Content delivery is revolutionised, appropriate language and currency will be displayed to the user. Sites are more able to comply with local advertising legislation.

3. Online advertising

At present 10% of online advertising is geographically based, yet in offline advertising this figure is 60% [6]. This is a massive area that is as yet untapped. Geographically targeted advertisement must be delivery in near real time and not affect the performance of a site.

4. Digital distribution

The Internet as a distribution medium will definitely increase. A geolocation service has a lot to offer web sites as it allows web sites to comply with local and international restrictions on digital content delivery by blocking unauthorised users from downloading content.

In this paper we describe our experiences in designing a methodology using existing Internet technologies, which provides a geolocation service that will let a given web site know in real time where a user is located. The main goals of the project were:

- High degree of accuracy: The service should produce a geographical location a high percentage of the time.
- Resolution: Precise targeting needed, to a country level.
- Global coverage: Service should cover all of the range of IPv4, which is approximately a 4.2 billon address space. It must also be extensible for when the next version of IP is released (IPv6).
- Distributed: By having the service distributed all the characteristics of distributed systems are included, such as reliability, scalability and strong enough for the largest enterprise applications

We begin by describing some of the technologies used to assist us in the geolocation service. We then describe the project design/methodology and discuss briefly the implementation. In the testing and evaluation we describe how we achieved the country to IP address mapping and finally we discuss the results and conclusions.

Internet Technologies

For the project the main technologies investigated were IP, Hyper Text Transfer Protocol (HTTP) and Domain Name System (DNS) with the emphasis on determining if they may be useful in the geolocation service.

IP addresses are what uniquely identify machines connected to the Internet; they are the basic unit of transport and the current version of IP is version 4 (IPv4). IP addressing was originally divided into a strict classful addressing scheme but this scheme has since been abandoned. IP addresses are local transparent and thus it was determined that IP addresses could not give any geographical location.

HTTP is an upper layer application that uses the TCP/IP protocol stack. HTTP is the standard protocol for communication between web browsers and web server. The main characteristics of HTTP as described by Comer [2] are that it is a request/response protocol and it is stateless. HTTP can be useful in getting the IP address of the machine connecting to the web server and it also includes in the `via` header the IP address of the proxy server between the browser and the server.

The Domain Name System is a distributed database. It allows local control of segments of the overall database, yet data in each segment are globally accessible through a client - server architecture. Programs called name servers make up half of the DNS client/server mechanism. Name servers contain information about segments of the database and make it available to clients called resolvers. Resolvers are usually library routines that query the name server.

The full domain name of any node in the tree is the sequence of labels from that node to the root. Domain names always read from node to the root (bottom up).

The Domain `dsg.cs.tcd.ie` contains four labels: `dsg`, `cs`, `tcd`, and `ie`. Any suffix of a domain name is also called a domain. Figure 1. contains four labels, at the lowest level is the domain name `dsg.cs.tcd.ie` (the Distributed Systems Group in the Computer Science Department in Trinity College), the second level domain is `cs.tcd.ie` (the domain name for the Computer Science Department in Trinity College), the third level domain is `tcd.ie` (the domain name for Trinity College), and at the top level domain is `ie` (the domain name for Ireland).

In practice domains often “collapse”, that is it could be under the control of another domain. It may be the case that the domains `dsg` and `cs` are in fact under the control of the domain `tcd`; then we say that `tcd` is a zone and will have a primary name server. All the Top Level Domains (TLD) are an authoritative zone for their domain and have a primary name server, they would usually also have secondary name servers.

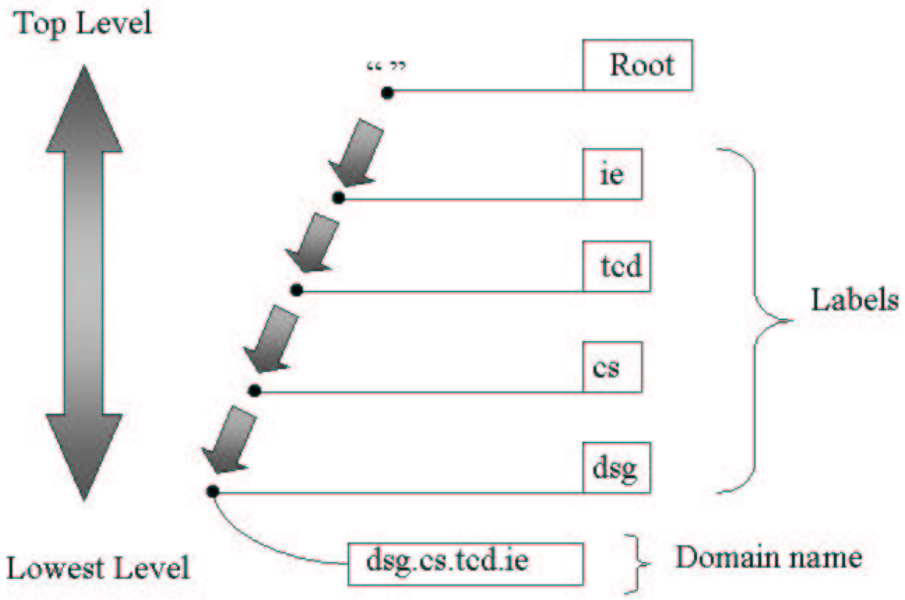

Figure 1. Operation of DNS

DNS requires that sibling nodes, nodes that are children of the same parent have different labels therefore this ensures uniqueness (i.e. the cs node can only have one child node with the label equal to `dsg`).

At the top level of DNS there is a rigorous structure, although below this there is not too much structure, there may exist a subdomain containing the letters A through Z (although they may strongly recommend against this). Originally the Internet top-level domain space was divided into seven domains:

com	Commercial organisations
edu	Universities and other educational institutions
gov	US government agencies
mil	US military organisations
net	Major network support centres
org	Organisations not mentioned above
int	International organisations

These are called the generic Top Level Domains (gTLD). To accommodate internationalisation of the Internet new top-level domains for every country in the world were developed and augmented to the system. Using ISO 3166 international standard for country codes; domain names were established for each country. (ISO 3166 contains the

official two letter abbreviations for every country in the world). All 7 gTLD and country domains are called the Top Level Domain (TLD).

DNS is a hierarchical naming scheme; it uses a distributed lookup in which domain server's map IP addresses to domains. Queries to the name servers are attempted to be resolved locally but if this fails the query must iterate through the tree.

The DNS name space is highly structured and works very efficiently at resolving IP addresses to domain names. But DNS was not design which geolocation in mind and thus it is difficult to retrieve geographical information from a domain name. We concluded that if an IP address resolved to a domain name that ended in a country code then we would infer that the user is located in that country.

In investigating the Internet technologies two useful utilities were discovered which helped in our geolocation service:

1. NSLookup is used to gather information about hosts in the network. It queries the Internet Domain Name Service (DNS) for specific information about hosts. It can be used with parameters:

   - [MX] Mail exchange records
   - [A] IP address
   - [CNAME] Canonical names

2. Whois is a database that keeps track of who maintains a domain name. All Internet domains must be registered with the certain authorities and each registered Internet domain is added to one of the authorities' database. This service is equivalent to a telephone book because it holds all information about the registered owner of the domain name including their geographical location.

   Whois servers:
   - For .com, .net, .org, .edu use "whois.networksolutions.com"
   - For .gov use "larc.nasa.gov"
   - For Asia/ Pacfic use "whois.apnic.net"
   - For Europe use "whois.ripe.net"
   - For various areas use "whois.ra.net"

   For the project a Whois proxy was used were requests were sent to the proxy and the proxy returned the relevant information.

Possible Approaches

Naïve approach
Probably most people when first asked how identify Internet user would come up with this solution. Based on the end of a domain name get the geographical location. By using

NSLookup domain names can be easily retrieved from IP addresses. But as stated this does not apply to some domain names (.com, .org).

More advanced
By using the Whois database the location of were an IP address was registered and is administered can be retrieved. The information returned by each authority varies so the information may not exist.

Narrowing the possible solutions
The World is broken up into different time zones; we decided that if we could get the time from the users machine we could then use this to narrow the possible countries a user may be located. The time has to be retrieved from the client's machine and not the HTTP header as the request may be going through a proxy server. A database stores the country – to – time zone mapping.

Project Design

The project methodology divides the top level domains into three different classifications based on their domain endings. By grouping the domains into this classification we can decide which domains give a geographical location.

The three distinct classifications:

Classification	**Domain Name Ending**
Type 1	.com, .net, .org
Type 2	.us, .edu, .mil, .gov
Type 3	ISO 3166 country code

Table 1.

If the domain ending is of Type 1 it was decided that it could not be used on its own and further processing would be necessary to determine a country. The Type 2 domain endings would represent the United States as it was deemed that most of these domains were issued to American institutions or bodies. The Type 3 domain endings represent the ISO 3166 countries, it was decided that these domain were issued to clients within the given country and they could be used to generate country identification.

The table below shows the actions to be performed based on the classification of the domain.

Classification	**Operation to Perform**
Type 1	1. Perform Whois query. 2. Check GMT database to narrow possible solutions. 3. Use the country returned from the Whois query.
Type 2	1. Use USA as the country.
Type3	1. Use the ISO 3166 country associated with the domain name ending.
Cannot be determined by DNS resolution	1. Perform Whois query. 2. Check GMT database to narrow possible solutions. 3. Use the country returned from the Whois query.

Table 2.

Type 1 domains are the hardest to identify in a geolocation service and need further processing. Type 1 domains also represent the largest percentage of the domain name space (approximately 62% or 68 million of the total domain name space).

Type 1 domains could not determined a location, but we decided that by querying the Whois database with an IP address or a domain name it returns information associated with the domain, and then by formatting and analysing this information returned a country can be obtained. By obtaining a time on the client machine we can also narrow the possible country choices.

Project Architecture

The project used the Model View Controller (MVC) paradigm, a servlet was used as the Controller, a JSP was used as the View and properties files represent the Model by containing localised text and image references. As the Controller also used session EJB‘s it spanned both the Web container and the EJB container.

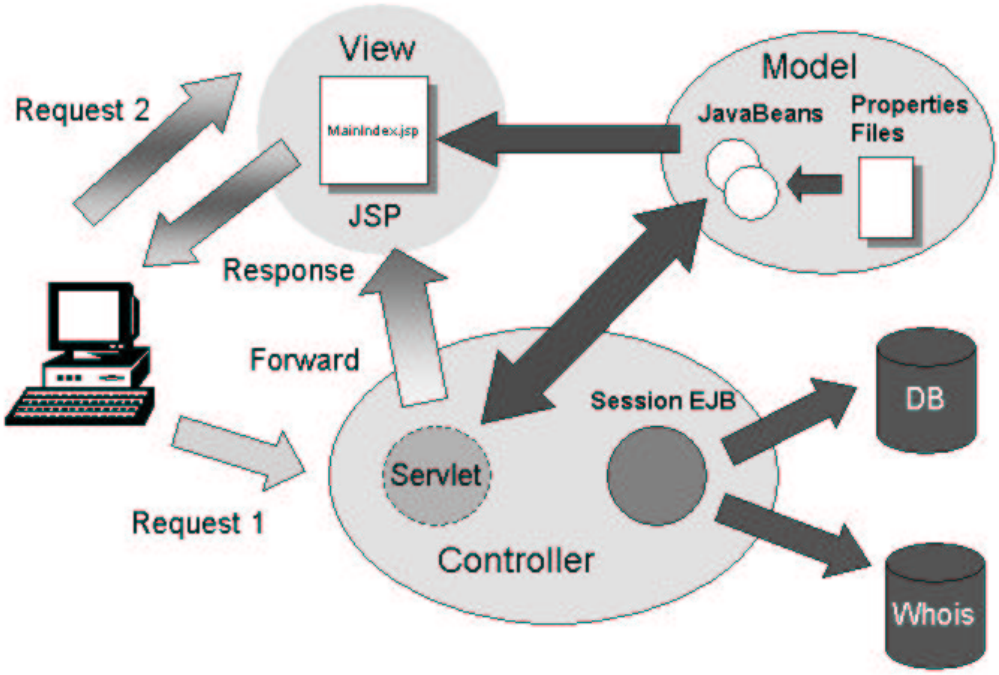

Figure 2. Project Architecture

Sun [5] recommend as well as using the MVC paradigm to divide the design into different tiers. For the project each tier had a specific role:

- Client tier: This tier contains the JSP page that displays the localised content to the user. It also obtains the time zone associated with the machine by using JavaScript and posting data back to the server via a hidden applet.
- Presentation tier: This tier used the same pattern as Sun's Front Controller pattern [5]. A single point of entry to the application is necessary where all requests are handled and dynamic content can be generated. This is referred to as the Controller.
- Business tier: A session bean is used to access the information stored in the resource tier. For the project the two resources were the GMT database and the Whois server.
- Integration tier: The Integration Tier uses JDBC to connection to the database and uses SQL commands to retrieve information from the database.
- Resource tier: This contains the GMT database were there is a country to time zone mapping; this is helpful in reducing the possibilities. It also contains the remote whois server.

Evaluation

It was decided that we would need a reliable country to IP address mapping to test the accuracy of the methodology. In order to get this mapping; random users on the Internet completed the questionnaire on the project homepage [7]. In total over 300 users completed the questionnaire.

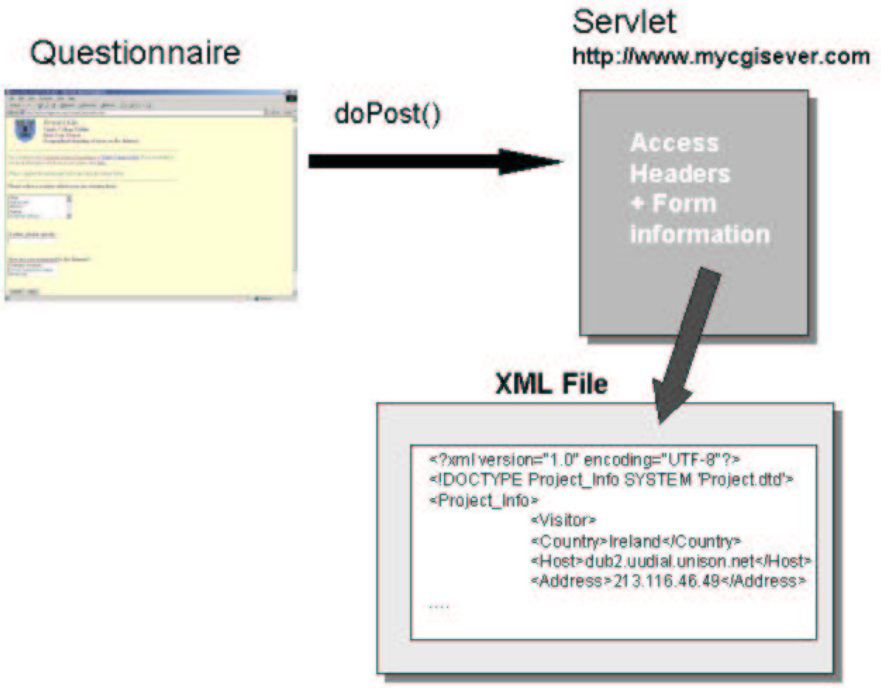

Figure 3. Implementation of questionnaire

Random Sampling

Chatfield [1] describes: "Statistics is concerned with collecting reliable numerical data and then analysing and interpreting them. … it should be emphasized that great care must be taken when collecting data, so that the selected *sample* is representative of the *population* from which it is drawn". This is very difficult when applied to the project, the *population* is every user on the Internet, and the *sample* is the 300 users that visited the web site.

Approximately 300 distinct users visited the site representing over 30 Countries. But how random is the random data? For the project an understanding of the distribution of the global domain names must be applied to the distribution of domains in the project data, this will verify if the data is stochastic.

Distribution of Domain Names

As discussed domain names resolve to IP addresses and vice versa. The domain name space is broken up into approximately 220 domains. For the project it was decided that if a domain name were of Type 2 then the country would be USA. But this is not necessarily true, there exists at least one university in Argentina that has a domain name ending in .edu (http://www.uncor.edu), and by the methodology the project would display a USA localised page, which is incorrect.

The major problem is that the distribution of Type 1 domain names are not allocated to given countries and that if a multi national company is allocated a block of IP addresses it can choose to distribute the IP addresses in any way it seems fit. This type of IP allocation is called non-commercial because they are not allocated to Internet Service Providers (ISPs) but companies and they are the hardest to identify geographical.

The domain name space consists of approximately:

- 68 million Type 1 domains
- 12 million Type 2 domains
- 30 million Type 3 domains

Figure 4 shows the domain name space breakdown for the data set. We can conclude that the data is random and is suitable for testing. But it must be stressed that more sample data is necessary for more in-depth statistical analysis.

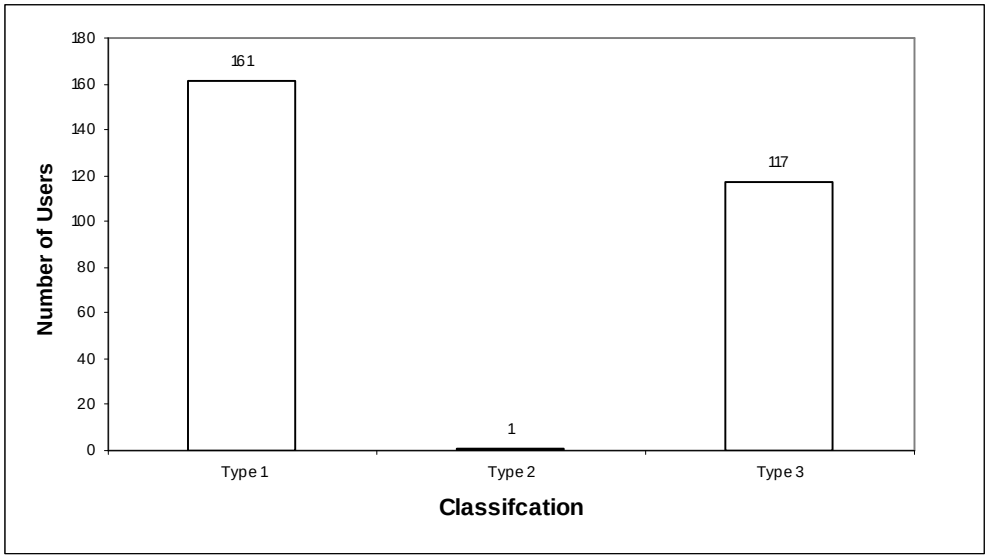

Figure 4. Number of users to web site

Results

Tests were performed using the data set collected from the web site. A correct result occurs when the country entered by a user on the web site is the same as what the methodology returns. An incorrect result occurs when the countries do not match.

The first results (figure 5) show the results using only NSLookup and the second (figure 6) show the results when using Whois; finally the third results show the combined approach is displayed.

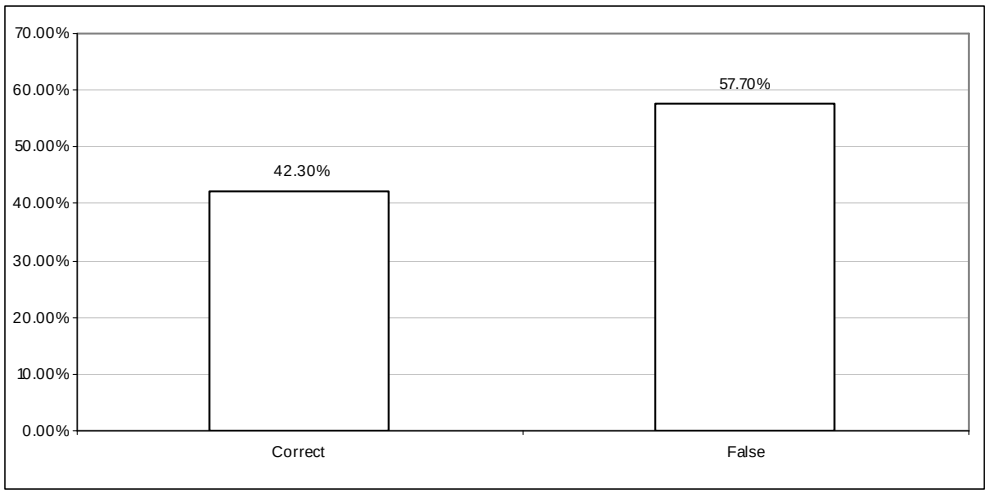

Figure 5. DNS Resolution results

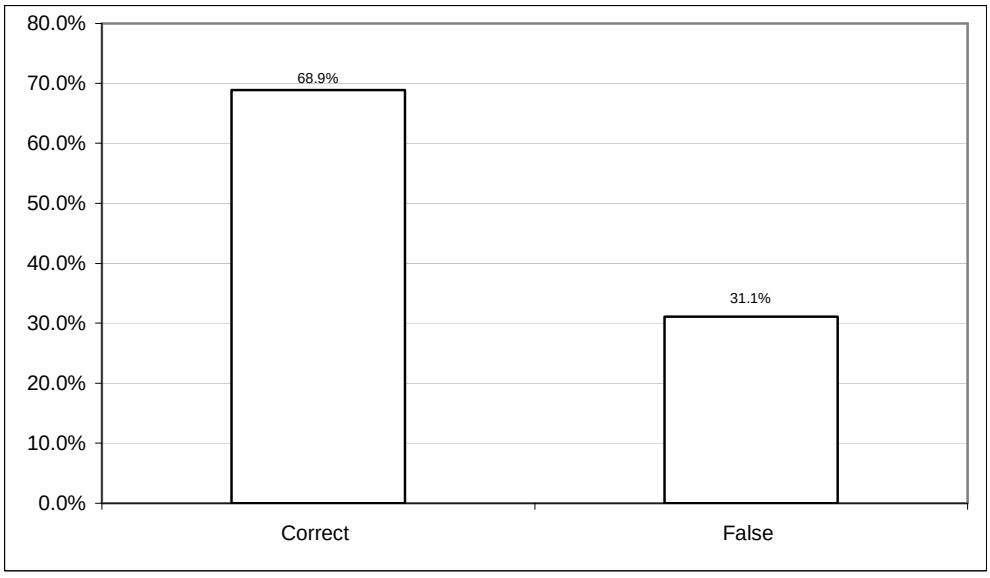

Figure 6. Whois server results

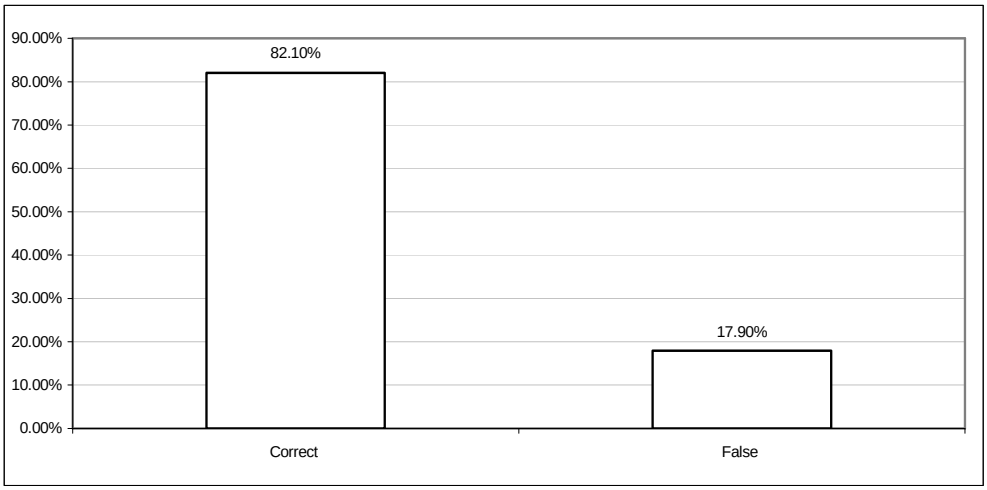

Figure 7. Combined approach results

The GMT database is used to narrow the possible countries a user may be viewing from; Figure 8 shows the number of countries per time zone.

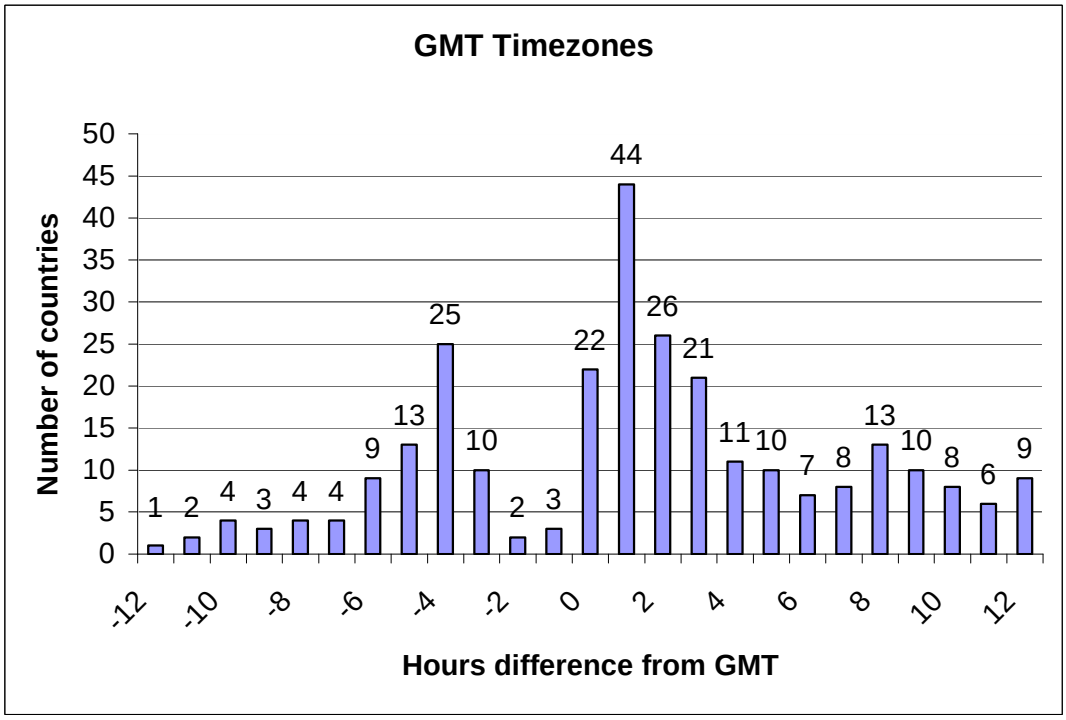

Figure 8. Distribution of countries based on time zone

The x-axis shows the time (in hours) difference from GMT (for example 10 equals Australia, and –5 New York (USA)). It may appear that there are more than 210 countries in the graph; this is because countries can span multiple time zones. The GMT results can narrow the possible countries from 220 (the whole world) to between 1 and 44.

Discussions and Conclusions

We have presented a methodology that gives a geographical location of a user in real time and with a high degree of accuracy. The methodology is lightweight in that it uses existing Internet technologies although not originally designed for geographical identification they can be manipulated to retrieve geographical information.

The project architecture ensured modularity; only one Enterprise Java Bean was used this ensured that it could be used with any J2EE compliant server. The main difficulties with the J2EE architecture is that there is a steep learning curve in development, but this is out weighed by the advantages it has to offer; reliability, robustness, scalability and the enterprise power.

Other approaches to the problem would involve mapping the full IP address space with an exact location. Companies such as Real Mapping and Qouva use this approach. Problems with this approach are that there is a lot of maintenance required as approximately 15,000 IP addresses change hand each month [6]. Also when the next version of IP is released IPv6; the address space will change from 2^{32} to 2^{128} addresses. This will make the maintenance of the mappings extremely difficult.

With the emergence of mobile devices such as laptops and palmtops users can now connect to the Internet at different locations as they migrate. Mobile IP allows a user to move freely from country to country using the same IP address. This would cause a problem for the methodology we have discussed and further research is necessary to see if anything can be done to remedy this problem.

The testing of the project consisted of a questionnaire; approximately 300 random users filled out the form. It was extremely important to have the IP to country mapping as otherwise it would have been impossible to verify the methodology. The accuracy of the methodology is not 100% accurate but it is a major improvement when compared to the naïve approach 42%, it is fact almost double at 82%.

The methodology is extensible and can accommodate IPv6. Each authority maintains their respective whois database this means the maintenance of application is taken care of by each authority.

Acknowledgements

This work was conducted as part of the author's undergraduate degree project at Trinity College.

References

1. Christopher Chatfield, Statistics for technology, Chapman and Hall, 1983
2. Douglas E. Comer, Internetworking with TCP/IP Principles, Protocols, and Architectures 4th Edition., Prentice Hall, 2000
3. CyberAtlas, Internet domain statistics, April 2001, http://cyberatlas.internet.com
4. Lisa Guernsey, Welcome to the World Web Wide. Passport Please, New York Times online, 15th March 2001, http://www.nytimes.com
5. Sun Microsystems, Sun Java Centre J2EE Patterns, First Public Reslease: Version 1.0, April 2001, http://developer.java.sun.com/developer/technicalArticals/J2EE/Patterns/
6. Qouva, In the press, Articles relating to Qouva, http://www.qouva.com
7. Project Homepage, http://www.mycgiserver.com/~howard_kim