2002

"Searching ……"

Karen Church

"Searching ……"

Karen Church[15]

The Institute of Technology, Blanchardstown
Karen.church@itb.ie

Introduction:

The Internet is a global collection of computer networks that in collaboration provides a powerful communications service and a comprehensive hierarchy of accessible information to millions of users worldwide. The Internet is used as an information retrieval mechanism by millions of different people across the globe. Searching for information on the Internet is an everyday occurrence for some people, but partially due to the volume of information stored within the Internet's infrastructure, it is not a trivial task. On submitting a request for information, a user can be overwhelmed by thousands of results that may have some relevance but may not be valuable to the user. The process of information retrieval on the Internet is a computing subject of interest to the general public.

History:

The Internet is probably the most revolutionary concept to enter the world of computers and communications. It enables the cost-effective interaction between users and their computers, it allows easy access to the largest source of information in the world and it brings broadcasting and multimedia capabilities never before perceived to the computing industry all without any geographic considerations [1].

The first origins of the Internet date back to 1962, when J.C.R. Licklider of MIT introduced his 'Galactic Network' concept. Licklider proposed the idea of a global interconnection of computers that could be used by people to access data and programs [1]. His vision is not unlike the Internet, as it exists today. In 1961 Leonard Kleinrock of MIT published a paper on the theory of packet switching in communications, which was a huge breakthrough in the field of computer networking. Kleinrock presented his idea to MIT researcher Lawrence G. Roberts who was working for DARPA, the Defence Advanced Research Projects Agency,

---

[15] This paper won the Higher Education and Training Awards Council Prize for Computing 2002

researching the field of computer networking. After agreeing that packet switching seem the way forward, Roberts continued working on the project and in 1967 he published a paper detailing his plans for ARPANET, the Advanced Research Projects Agency [1].

The entire field of networking, connections, Internet protocols and communications was researched over the following years by ARPANET. From one node in 1969 to 23 by 1971, the numbers of nodes connected by ARPANET broke 1000 in 1984, 100,000 in 1989, 1,000,000 by 1992 and over 50 million by 1999. Throughout the years APRANET were responsible for a number of networking breakthroughs including the creation of the TCP/IP protocol [1]. While research and development was being carried out by scientists in the field of computer networking, a new concept called the World Wide Web was also being conquered.

Tim Berners-Lee, a CERN computer scientist, introduced the World Wide Web in 1989. The World Wide Web (WWW) is a system that essentially sits on top of the Internet and provides an information sharing mechanism. The idea behind the WWW was to provide users with a simple, single user interface written in hypertext that would enable users to browse the information on the Internet and to access that information in an easy and efficient manner [2].

The original proposal of the WWW anticipated a global hypertext system known as HTML, which had been developed by Tim Berners-Lee. This new language would enable web documents to be indexed and formatted for presentation over the Internet. A simple protocol that would process user requests for documents and present them to the user was also proposed called HTTP, the Hypertext Transport Protocol. During the course of the 1990s the concept of the WWW and the technologies behind it blossomed and the growth of the Internet and the WWW exploded to depths that were inconceivable by the mid 1990's [2].

Today the Internet and the WWW are considered a necessity for users and businesses alike. They provide companies with the ability to establish a global presence and to remain accessible to customers at all times. They enable companies to partake in eCommerce and eBusiness where business processes, marketing ventures, business transactions and customer service facilities all take place online. Users are able to interact and communicate with other people in a very cost-effective way, shop online using interactive shopping cart applications and secure payment systems, and find an abundance of information on almost any subject area.

Searching the Web:

For users to find information, products or services online, they generally employ the services of a search engine. Search engines work on behalf of a user, by receiving a query as input and carrying out a search of the Internet, in order to find relevant information for the user. Users frequently encounter problems while searching for items on the Internet for a number of reasons. The first reason is that the sheer volume of information present on the Internet makes trying to find a specific item increasingly difficult and although search engines are very useful tools for searching the Internet, they can be very inefficient. All who use the web have had experiences with search engines where we enter our keywords, click on submit and receive either hundreds of thousands of results or very few. When we do receive too many results, we are then presented with the exhaustive task of trawling through those results in order to find what we were actually looking for. Although the web and search engines provide efficient access to an increasing amount of information, there are definite limitations to current search engines and the existing method of information retrieval.

Search engines are search tools that assist users in finding items on the Internet. Search engines use software agents that survey the Web and build databases of indexed web documents that are then used to carry out searches on behalf of the user. There are a number of different types of search engines in existence but every search engine performs three basic functions [4]:

- They search the Internet for special words found in web pages.
- They store an index of these words along with the details of where to find the websites that contained the words.
- They allow users to look for words or combinations of words found in the indexes that they build.

Today search engines index hundreds of millions of pages, and respond to tens of millions of queries per day [4]. To find information on the hundreds of millions of web pages that exist, a search engine uses the services of software robots, called *spiders*. Spiders search the Internet and build lists of the words they find on web sites. This process is called *web crawling*. Most web pages have Meta tags built into them. Meta tags are keywords entered by the owner of the web page that influence the way in which a web page is indexed [4].

The data obtained by the spiders has to then be indexed and stored with additional information that makes it useful to the search engine. Normally this additional information

includes ranking and weight information that dictate the order in which web pages are returned as a result to a user query. Once the search engine has designed and built a useful index, users are then able to submit queries to these indexes in order to find the information they desire [4].

Despite the fact that search engines provide substantial assistance in locating and presenting information to users, a number of problems can occur quite frequently. The first problem is that each search works in a very different manner, covering a different domain of web sites, and requiring different syntax for queries. For example one search could require you to enclose a query in quotation marks, while another search engine might require you to use Boolean operators such as AND and OR to refine you query. It is very apparent that to carry out a keyword search in the most efficient manner, a user must have a comprehensive understanding of the mechanics behind search engines, which is something that the majority of Internet users do not have [5].

The second problem with search engines is regarding the results you receive after submitting a query. In general users either receive too many results or two few results. Most search engines order the search results using ranking information. A number of factors influence the ranking of websites. These include Meta tags, web page content, the format of your HTML, etc which although affect the ranking of a website, have nothing to do with the value of the information within the web pages. The accuracy of a users search is also influenced by the number of keywords you enter, which makes very little sense when it means that if a user enters lots of keywords to ensure precision, they are inevitably and unconsciously reducing the relevance of their search [5].

Another issue, and probably the most exasperating problem is that once you complete a search and finally find what you are looking for, you can't easily return after a number of days or weeks to the same result set, unless you remember the search engine you used and the keywords you entered or you remembered to book mark the web page [5]. Ideally every user should be able to carry out an Internet search that returns personalized, relevant and valuable results in an understandable format and to which a user can return when ever they wish.

Time for Change:

What if a system existed, that not only enabled the user to carry out searches that returned relevant and valuable results, but it did it in a manner that removed the current search effort away from the user? What if the traditional mechanism of user and business interaction was completely changed? In the traditional mechanism a businesses will build an interactive website that carries descriptive examples of the product ranges and services the business has to offer. The business will then deploy the website online, partake in some advertising and search engine submission procedures and essentially wait for customers or users to find them. Customers, on the other hand have to carry out all the searching in order to find information about a particular business. Imagine a system in which the current roles were reversed. A system that reduced the present search effort of the customer, and placed emphasis on the business carrying out the work in order to access more customers. This suggested system would involve replacing the traditional information retrieval mechanism with an improved system that allows the customer to simply state what they want, sit still and wait to be found while the business carries out the searching.

I would like to suggest a system that could be used to change the way the Internet currently operates in order to address the various information retrieval issues I outlined previously. The system would be able to reduce the tedious search process that customers have to participate in and place onus on the business to find customers using two new and exciting technologies to provide its functionality, namely Intelligent Agents and Web Services.

According to Professor Michael Wooldridge, a pioneer in intelligent agent systems, an Intelligent Agent is defined as [6]:

> *"a computer system that is capable of flexible autonomous action in order to meet its design objectives."*

Autonomous is defined in the English dictionary as independent, self-sufficient, self-ruling and self-directed. For an agent to be autonomous it means that it can operate efficiently without intervention or interruption from humans or from other agents. To be flexible, an intelligent agent must present a number of key capabilities. The capabilities listed below are adapted from three sources [6], [7], [8]:

- **Responsive** – intelligent agents should be able to *perceive* or sense their environment. The intelligent agents should also be able react in a sensible and well-timed manner to changes that occur in their environment.

- **Proactive** – intelligent agents should be able to display goal-oriented behaviour. This means that the agents should not simply respond to their environment, they should *take initiative* when and where it is appropriate.

- **Social** – an intelligent agent should be able to communicate, collaborate, and interact with other intelligent agents and humans in order to solve problems and to meet their design objectives.

Web Services have emerged as a new and exciting technology destined to revolutionise the way in which businesses operate over the Internet. Using web services technologies; businesses can publish their business processes, locate and subscribe to other web services, and exchange information in a very efficient manner.

The Stencil Group, a research and analysis firm that specialises in web services technologies define a web services as [9]:

> *"Loosely coupled, reusable software components that semantically encapsulate discrete functionality and are distributed and programmatically accessible over standard Internet protocols."*

The definition above specifies that the first attribute of web services as loosely coupled. More traditional applications depend on the tight interconnection of all the auxiliary elements. Tightly coupled systems require developers to thoroughly understand and control both ends of the connection between the elements. Tightly coupled systems also make it very difficult to extract one element and replace it with another element. Loosely coupled systems require a simpler level of management and enable flexible and easy reconfiguration of component elements [9].

The second attribute is that web services are reusable software components. Web services allow developers to reuse code that has been created by other developers by allowing the developers to assemble and extend the code blocks in new ways. The third attribute, to semantically encapsulate discrete functionality, means that web services are self-contained components. By providing information about the inputs and outputs they expect, other

software is able to invoke its functionality. Programmatically accessible means that web services were originally designed to operate at code level, meaning that they are not designed for direct human intervention. They enable software-to-software interaction, although they can be programmed to accept human interaction. They are distributed over the Internet and make use of existing standard Internet protocols such as HTTP. By leveraging existing Internet protocols, web services provide a standard component-based architecture [9].

In this suggested system the customer should be able to find the items they are looking for with minimal effort and businesses should be able to access more appropriate customers in a more efficient manner. The suggested system consists of three major components – a Customer Intelligent Agent (CIA), a Business Intelligent Agent (BIA) and an Agent Portal (AP), that in combination will enable a customer to make a request and receive what they actually want in a timely fashion, and will empower businesses with the tools to carrying out searching in order to find new customers. The system will allow businesses to register themselves as available and customers will be allowed to register themselves as seeking service. The power in the system will not be in the availability of customer data and supplier data but in the manner in which it can be matched and retrieved, giving best fit to both business and customer, easing the use of the web and enhancing the overall customer service.

The Customer Intelligent Agent (CIA) can be thought of as a type of personal assistant or intelligent servant that will work on behalf of the customer. The CIA presents the customer with timely and relevant search results with minimal intervention or interaction from the customer. When a customer wants to find something on the Internet he/she will input a request to a CIA. The CIA will take the request from the customer, process and structure the request in order to extract the real meaning and post the request in its new format to the Agent Portal (AP). The goal of the CIA is to quickly find what the customer actually wants.

The Business Intelligent Agent (BIA) can be thought of as the businesses own cost-effective, reliable and flexible sales person that will work on behalf of the business. When a business wants to find customers, it inputs specific information to the BIA including details on the products and services the business has to offer, the location of the business, the type of customers that the business wants to attract, etc. The BIA will then process and structure that information into a specific format so that it may use this information to find customers for the business. In order to find customers the BIA searches all customer requests that are posted in the Agent Portal (AP). The BIA scans the Agent Portal (AP) looking for customer requests it thinks it can handle. If the BIA finds a request it thinks its business can satisfy it makes contact with the CIA holding that request and the two agents begin to communicate in order

to conclude as to whether or not the business can satisfy the customer's request. If the request can be satisfied, both the customer and the business are notified.

The Agent Portal (AP) can be thought of as a multi-agent web-based system in which CIA's and BIA's communicate and collaborate in order to satisfy their goals. The AP acts like a type of bulletin board where CIA's post customer requests. BIA'S scan the AP searching for customer requests that it thinks its business can satisfy.

Both the CIA's and BIA's would have to exhibit intelligent, flexible, social and adaptable behaviour. They would be expected to make decisions on behalf of their owner's. For BIA's this means deciding whether its business could handle particular customer requests or not. For CIA's this means that when they receive notification from a BIA that a request may be satisfied, the CIA must decide whether or not to accept the offer made by the BIA. Both agents must also interact with their owners and keep their owners informed of any developments as they occur. These Intelligent Agents would be exposed as web services, allowing them to be flexible, open and interoperable. Using web services to expose the Intelligent Agents allows the system to take advantage of the open, dynamic and platform independent features of web services, as well as the aptitude and flexibility of Intelligent Agents. In order for an intelligent agent to be mobile, the agents need to share a common platform with its colleagues. The best candidate for providing a standard platform would be web services technologies.

At present, no such system exists that uses Intelligent Agent and Web Services technology to reduce the search effort from customers, to place responsible with businesses to carry out the searching and for customers to simply make their request and wait. The closest types of systems do involve the use of agent technology. They take the form of software agents that remove some of the tedious tasks associated with searching the Internet away from user. The majority of these systems simply take a search query from a user and simultaneously submit the search to multiple search engines, and then apply some filtering to the results. Although such software systems remove some of the search effort away from the user, the results are still not relevant enough. The agents are still using existing search engines to satisfy user requests and although what the user is looking for may be on the Internet, it may not be ranked high enough amongst the top search engines to even be included in the search agent's results.

Conclusion:

In my discussion of the Internet and its use to customers as an efficient information retrieval mechanism, I have posed the problems and inefficiencies that exist with the current search mechanism. I have provided an overview of search engine technology as a tool that facilitates users in finding information on the Internet and I have reviewed the problems with current search engines. I introduced a new system that could change the current operation of the Internet by reducing the current search effort made by the customer and by placing the responsibility of searching on the business. In using such a system the customer should be able to find the items they are looking for with minimal effort and businesses should be able to access more appropriate customers in a more efficient manner. The suggested system would provide its functionality by combining two exciting technologies, namely Intelligent Agents and Web Services.

References:

1 "*A Brief History of the Internet*", Barry M. Leiner, Vinton G. Cerf, David D. Clark, Robert E. Kahn, Leonard Kleinrock, Daniel C. Lynch, Jon Postel, Larry G. Roberts, Stephen Wolff, See: http://www.isoc.org/internet/history/brief.shtml
2 "*History of the World Wide Web*", See: http://www.wdvl.com/Internet/History/
3 "*A Helpful Guide to Web Search Engines*", See: http://www.monash.com/spidap4.html
4 "*How Internet Search Engines Work*", See: http://www.howstuffworks.com/search-engine.htm
5 "*Why search engines are clueless*", See:
   http://mappa.mundi.net/inform/archive/inform_0169.html
6 *Wooldridge, M. (1996), "*An introduction to multi-agent systems*"
7 *Wooldridge, M. and Jennings, N. R. (1998), "*Applications of Intelligent Agents*"
8 *Wooldridge, M. and Jennings, N. R. (1995), "*Intelligent Agents: Theory and Practice*", The Knowledge Engineering Review*
9 *The Stencil Group web services definition. See
   http://www.stencilgroup.com/ideas_scope_200106wsdefined.html*