INTELLIGENT INFORMATION AGENTS: SEARCH ENGINES OF THE FUTURE?

Naomi Augar [1]

[1] School of Information Technology, Deakin University

ABSTRACT: Information gathering on the Internet is a time consuming and somewhat tedious experience. The main method for gathering information on the Internet at present is the search engine. Literature reviewed indicates that Intelligent Information Agents can improve the process of gathering information on the Internet. However, such agents are not widely used at present. This research reviews several Intelligent Information Agents to determine whether they are a viable alternative to search engines.

INTRODUCING AGENTS AND RELATED LITERATURE

What is an Agent?

Software processes that act on behalf of the user are known as *agents*. All agents exhibit similar characteristics, regardless of whether they are people or a piece of software. Campbell [Cam99] describes an agent as a process that performs tasks on behalf of the user, by applying its specialised knowledge. It makes decisions about how to complete the tasks, and has the ability to learn the preferences of the user, to improve its performance in the future.

Franklin and Graesser [FG96] define an agent as an autonomous process running on a computer that is able to sense and react to its environment. As an autonomous process, an agent is able to run without interaction with the user and must therefore be able to make decisions about the environment and the realisation of its goals. Hermans [Her96], Jennings and Wooldridge [JW96] outline several characteristics of software agents. They believe agents require social ability to interact with the user. The agent must be responsive and proactive so it can sense and react to its environment and the users needs. They must be temporally continuous, goal oriented and adaptive. Finally, agents should be autonomous and able to collaborate with the user and other agents to perform tasks.

An *Intelligent Agent* is an agent that uses stored knowledge, related to its tasks and user preferences, to aid in its performance of tasks and the achievement of its goals. An *Intelligent Information Agent* is an Intelligent Agent that locates, collates and manipulates information contained in stored resources on a distributed information network [CG01]. Intelligent Information Agents communicate with the user via their interface. The user also views results and information provided by the Intelligent Information Agent through the agent interface.

Haverkamp and Gauch [HG98] and Heilmann [HK95] contend that Intelligent Information Agents should have processing power, knowledge of their environment and a domain and information model. Processing power is the ability to decompose a user's query into sub-queries, interpret the results, and provide additional processing when necessary. Knowledge of its environment implies that the Intelligent Information Agent must have an understanding of the resources at its disposal, and how to gain access to these resources. An Intelligent Information Agent should also have domain and information models which allow it to infer the context of a given query. These models allow it to locate relevant information for the user even if the query supplied does not sufficiently describe the user's information need.

Why use Agents?

Negroponte [Neg97] believes that agents are useful not because they can perform tasks a user could not perform on their own using other tools, but because they perform tasks, the user finds trivial or mundane. By delegating the task of information retrieval to the agent, the user is able to direct their attention to tasks that are more enjoyable or make better use of their time.

Lieberman [Lie97] believes that autonomy, the ability to act on ones own initiative, and temporal continuity, the ability to run without pause, are two kinds of functionality that make agents useful. The agent may discover information of value to the user and notify them. It can remain active for any period required to fulfil requests, even if the user has logged out. The ability for the agent to run whilst the user directs their attentions to other matters allows the user to truly delegate tasks to the agent.

Negroponte [Neg97] suggests that increasing the amount of information that the user has access to on the Internet does not improve the Internet as an information resource. Rather, it makes the process of finding accurate reliable resources that match the user's information need even more difficult. Negroponte [Neg97] believes that the user now has access to far more information than they can possibly absorb. He contends that users do not need more information. They need a relatively small amount of information that is concise, accurate and relevant to them. Therefore, a method of filtering the abundant information, so that only poignant information remains, is required.

The main function of an Intelligent Information Agent is locating information resources for the user. Different agents use different methods for managing information. An agent's competence ultimately depends on its ability to satisfy the information needs of its user. As such, its ability to retrieve the right amount of quality information quickly is important.

The agents considered in this paper act on a single user's behalf and run on that user's personal computer. As such, it is not feasible to create an agent that requires vast amounts of disk space and system resources to run. If an agent creates its own index of Internet sites, as search engines do, it would consume more disk space than a personal computer has available [BR99]. Rather than maintaining their own index of Internet sites, agents can develop complex queries that they send to search engines that maintain and utilise their own comprehensive Internet indexes. This gives the agent the benefit of an index without having the problems associated with storing and managing an index. Some agents use other methods for locating information on the Internet [LF01]. Different methods are illustrated by the sample agents discussed in later sections of this paper.

Information management in agents is not restricted to the management of user queries. Agents also use a knowledge base to help them manage information, and create a model of the user's information needs. Maes [Mae97], identifies an approach to the design of agents that is "knowledge-based". It involves the agent accruing its knowledge base over time. Maes [Mae97], uses a machine learning approach to the design of agents. She notes that an agent can "learn" in several ways. Firstly, the agent may monitor user behaviour and actions, with the aim of detecting patterns that it can emulate and automate. Indirect or direct user feedback allows the agent to acquire competence. Explicit examples given by the user can also train the agent. Finally, the agent may seek advice from other sources or agents that provide the same service to their user and have more experience.

An agent that learns or accrues knowledge exhibits adaptive functionality. It notices things and tries to detect important events or information. It recognizes and interprets events using a set of rules and responds to events using a set of action rules [Eri97]. This gradual learning process means the agent becomes more competent with time [Mae97]. This upward trend in competence also eases the user into trusting the agent, and delegating tasks to it. As the user sees the agent become more helpful and competent, they are more able to trust it.

Literature indicates that agents are a useful alternative to search engines because the user can delegate the information retrieval task to them. Autonomy, temporal continuity and adaptiveness are some of the aspects of agent functionality that enable an agent to work effectively on behalf of a user. The following section reviews several existing agents to highlight their flaws and strengths. This review aims to test the theory that agents are a viable alternative to search engines.

RESEARCH OVERVIEW: AN AGENT SURVEY

Four agents were studied as part of this research, *Alexa*, *Copernic2001*, *LexiBot* and *WebMate*. These agents were free for download from the Internet. The functionality and design of each of the sample agents was reviewed using a survey template to ensure that all agents received a consistent evaluation. Sample queries were developed so that each agent could be repetitively used so as to test and analyse each agent's process of information retrieval compared to that of a search engine. Table 1 summarises the information retrieval process for each of the surveyed agents. It is followed by a brief summary of the survey results for each sample agent.

Table 1. Decomposed Information Retrieval Process for the Agent Sample

Alexa	***Copernic2001***	***LexiBot***	***WebMate***
Extract URL	Accept Query	Accept Query	Lexically Analyse Site Browsed by User When Prompted
Send URL query to Central *Alexa* Database of User Browsing Patterns or Send Query to Multiple Search Engines	Send Query to Multiple Search Engines	Send Query to Multiple Search Engines	Extract Keywords Create Keyword Profile
Receive Results	Receive Results	Receive Results	Send Complex Query of Keywords to Search Engine /Lexically Analyse Resource Sites as Prompted
Present Results	Filter Results using Algorithm	Filter Results using Algorithm	Receive Results
	Present Results	Present Results	Present Results

Alexa

Alexa has a toolbar like interface that appears within an Internet browser window. *Alexa* provides unsolicited recommendations that appear as hyperlinks within the toolbar. *Alexa's* recommendations are autonomous. The related hyperlinks that Alexa provides are determined by observing the browsing habits of other *Alexa* users [Ale02]. Recommendations are based on what other *Alexa* users, who have viewed the site currently being browsed, have browsed in the past. No lexical analysis or keyword profile is used to produce the recommendations. They are purely peer reviewed. The user can also define their information need as a query entered via the search box provided in the *Alexa* Toolbar. The query may have either a Boolean or a simple English format.

Copernic2001 Basic Edition

Copernic2001 basic edition did not act autonomously to locate information resources for the user. As such, it did not adhere to the definition of an Intelligent Information Agent as defined earlier in this paper. However, it acts as an information filter. *Copernic2001* has a search box where the user can enter their information need, or query. Queries may be expressed in a Boolean format, or in a simple question or simple English format. *Copernic2001* conducts searches by sending user queries to multiple search engines.

LexiBot

LexiBot does not act autonomously to locate information resources for the user. Its functionality is similar to a meta-search engine in that it compiles and collates query results from several search engines and presents them to the user in an order of relevance to the query. As such, its functionality and information gathering process model is identical to that of *Copernic2001*.

WebMate

WebMate cannot be used to find results for explicit queries. Rather *WebMate* accrues information about user preferences when the user evaluates the page currently being browsed. When the user clicks on an icon that *WebMate* places on each page, *WebMate* adds information to their user profile. WebMate breaks the page the user recommends into keywords and calculates a "weight" for each keyword added to the user profile. *WebMate* uses lexical analysis to extract keywords with high frequencies from browsed pages to send to search engines to provide recommendations [CS97].

WebMate provides recommendations to the user in two ways. The first involves sending queries to several search engines. The queries are formed using a keyword profile of the user's information needs. *WebMate* develops this profile by extracting high instance keywords from the web pages when prompted by the user. *WebMate* places a clickable icon into the HTML of web pages viewed in the browser. The user can click this icon to show their approval for a given page. This approval prompts the lexical analysis of the page to begin, and the keyword profile to be updated. The second method, involves searching sites contained in *WebMate's* resource list for articles that contain terms in the keyword user profile. Once a user profile is accrued, the user can prompt *WebMate* to recommend sites to them.

Measuring Information Retrieval Performance

The performance of each agent in terms of information retrieval was compared to that of search engines by defining an ideal result set for a specified query and using the performance measures of precision and recall recommended by [BR99]. These measures can be used to produce a precision versus recall curve that depicts the information retrieval performance of each agent.

The ideal result set was established using several major Internet search and meta-search engines. MSN.com and Google were selected based on their popularity, as established by Nielsen's NetRating Survey for June 2002 [Sul02]. MSN.com was the most popular search engine in June of 2002, receiving 28.6 per cent of Internet user queries [Sul02]. Google received 26.4 per cent of Internet user queries. To ensure that a variety of possible results was collected a less popular search engine was also selected, AltaVista. AltaVista received 4.7 per cent of Internet user queries in June 2002 [Sul02].

Each search engine was queried to find information relating to anthropomorphic information agents. A series of Boolean queries for this topic were devised and posed to each search engine and the results of every search logged. Every web page was reviewed and its relevance to the topic was determined. For a page to be relevant it had to contain all three terms, "anthropomorphic", "information", "agents", and they had to appear in the context of Information Technology. The results produced were then merged, to remove duplications, and the ideal result set, summarised in Table 2 was produced.

Table 2. Ideal Result Set

Total Distinct Pages	Total Distinct Irrelevant	Total Distinct Relevant
91	56	35

The Internet contains a total of *N* websites (represented by the ideal result set) that are relevant to the user's information need. When a query is evaluated, a total of *Q* websites will be returned to the user, of which Q_r will be relevant and Q_i will be irrelevant or unreachable. Hence:

$$Q = Qr + Qi$$

Recall is the portion of relevant web sites retrieved from the total set of relevant web sites available.

$$\text{Recall} = \frac{Qr}{N}$$

Precision is the proportion of retrieved material that is actually relevant.

$$\text{Precision} = \frac{Qi}{Q}$$

Information retrieval performance was evaluated by posing the query '"information agents" AND anthropomorphic' to each agent where query processing functionality was present (variants of this query were also posed; however, only one, considered representative of the overall results is

discussed here). The precision versus recall curve for the two agents that had working query processing facilities is depicted below in Figure 1. Only *Alexa* and *Copernic2001* produced a result set for this query. The result set produced by *Alexa* was identical to the Google result set which was generated whilst creating the ideal result set. Although *LexiBot* provided query search access, no results were returned for the query. *WebMate* does not accept queries, as explained previously, relevant documents must be browsed and recommendations are then supplied on demand. To determine *WebMate's* information retrieval performance, it was used to browse documents in the ideal result set with the aim of building up a relevant user profile, from which it could derive its recommendations. Unfortunately, *WebMate* failed to recommend a single relevant document from the ideal result set after the profile had been established.

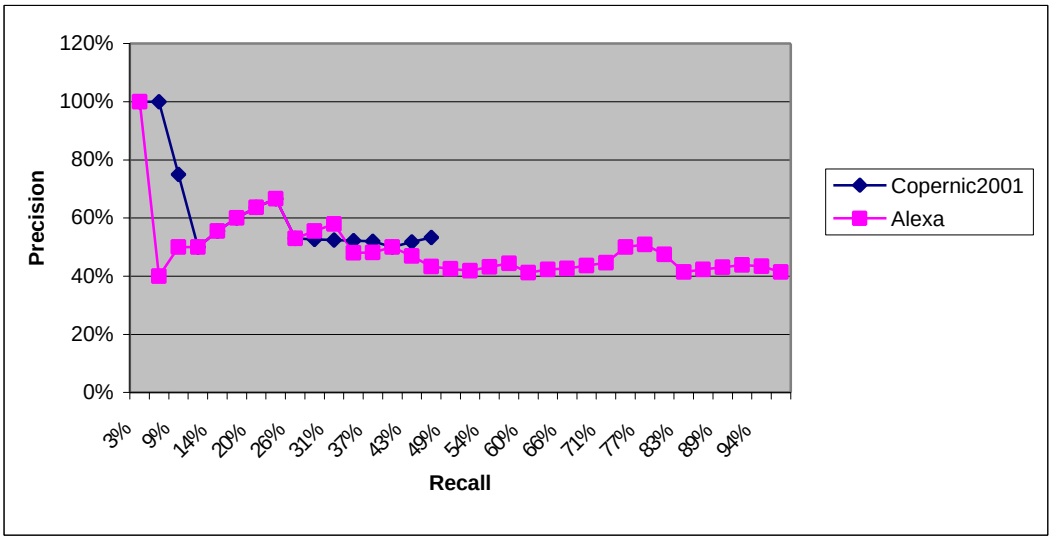

Figure 1. Precision versus Recall Curve for Alexa and Copernic2001

The relative results for recall versus precision, for both *Alexa* and *Copernic2001,* are compared in Figure 1. The graph shows that both agents follow a similar pattern in terms of recall and precision, up to levels of approximately 50 per cent recall. At this point *Copernic2001* produced no more results. *Alexa* went on to locate all but one of the results in the ideal result set, achieving 97 per cent recall. Precision is initially high but levels off at an average of fifty per cent. Maintaining levels of recall and precision at a constant 100 per cent is ideal for an agent. However, it is unlikely that a given agent or search engine will consistently produce a result set containing only documents that are relevant to the user.

CONCLUSION

None of the four sample agents strictly conforms to the definition of an autonomous, temporally continuous agent, defined in earlier sections of this paper. Two of the sample agents act as information filters rather than agents. The information retrieval performance of the sample agents failed to exceed that of the search engines, as such there is no real benefit, in terms of the quality of information retrieved, to using the agent instead of a search engine. In fact the user may only elect to use one of the sample agents because of a preference in interface layout, as there is no real technical superiority compared to a search engine.

Negroponte [Neg97], points out that agents are useful, not because they can perform tasks that people can perform already, but because people can delegate tasks to them. Only two of the agents provided alternatives to a query based system that mimics a search engine. Both alternatives were unsuccessful in providing relevant results. It was not possible to delegate the information retrieval process to any of the sample agents.

Whilst many intelligent information agents exist, few are available free for download. Whilst literature touts intelligent information agents as a time saving, efficient alternative to search engines, the agents

reviewed here are little more than interfaces to existing search engine technology. Until functionality such as autonomy and temporal continuity is realised in free, easily accessible and usable agents, search engines will remain the dominant technology for information retrieval on the Internet.

REFERENCES

[Ale02] Alexa_Internet_Inc (2002). *Alexa WebSearch - Quick Tour*, Alexa_Internet_Inc, Internet, [accessed: 22 July 2002], <http://pages.alexa.com/prod_serv/quicktour_new.html>.

[BR99] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern information retrieval*. Essex, ACM press.

[Cam99] Campbell, M. (1999). *Information Discovery Using Intelligent Agents*. Technical Report, TR C99/17, Deakin University. Melbourne.

[CG01] Campbell, M. and Gozinski, A. (2001). *An Information Agent Model*. Technical Report, TR C01/19, Deakin University. Melbourne.

[CS97] Chen, L. and Sycara, K. (1997). *WebMate:Browsing Applet*, Chen and Sycara, Help File, [accessed: 24 June 2002].

[Eri97] Erikson, T. (1997). *Designing Agents as if People Mattered*. Software Agents. J. Bradshaw. AAAI and MIT Press, Menlo Park.

[FG96] Franklin, S. and Graesser, A. (1996). *Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents*. Third International Workshop on Agent Theories, Architectures, and Languages, Springer-Verlag.

[Her96] Hermans, B. (1996). *Intelligent software agents on the internet: An inventory of currently offered functionality in the information society and a prediction of (near-) future developments*. Ph.D. Thesis, Tilburg University. Tilburg.

[HG98] Haverkamp, D. and Gauch, S. (1998). *Intelligent Information Agents: Review and Challenges for Distributed Information Sources*. Journal of the American Society of Information Science 49:4(April): 304-311.

[HK95] Heilmann, K., Kihanya, D., et al. (1995). *Intelligent agents: A technology and business application analysis*, Intelligencia Inc., Internet, [accessed: 26 October 2002], <http://www.mines.u-nancy.fr/~gueniffe/CoursEMN/I31/heilmann/heilmann.html#exsum>.

[JW96] Jennings, N. and Wooldridge, M. (1996). *Software Agents*. IEE Review(January): 17-20.

[LF01] Lieberman, H., Fry, C., et al. (2001). *Why Surf Alone?: Exploring the Web with Reconnaissance Agents*, Internet, [accessed: 20 February 2002], <http://lieber.www.media.mit.edu/people/lieber/Lieberary/Letizia/Why-Surf/Why-Surf.html>.

[Lie97] Lieberman, H. (1997). *Autonomous Interface Agents*, Henry Lieberman, Internet, [accessed: 5 February 2002], <http://lieber.www.media.mit.edu/people/lieber/Lieberary/Letizia/AIA/AIA.html>.

[Mae97] Maes, P. (1997). *Agents that Reduce Work and Information Overload*. Software Agents. J. Bradshaw. AAAI, Menlo Park.

[Neg97] Negroponte, N. (1997). *Agents: From Direct Manipulation to Delegation*. Software Agents. J. Bradshaw. AAAI and MIT Press, Menlo Park.

[Sul02] Sullivan, D. (2002). *Nielsens Net Ratings*, Sullivan, D, Internet, [accessed: 20 August 2002], <http://searchenginewatch.com/reports/netratings.html>.

Alexa: <http://download.alexa.com/alexa7/startpage.html?p=404_redirect>.

Copernic2001: <http://www.copernic.com/en/downloads/index.html>.

LexiBot: <http://www.LexiBot.com/>.

WebMate: <http://www-2.cs.cmu.edu/~softagents/WebMate/download.html>.