Information Technology Journal 3 (2): 196-201, 2004
ISSN 1682-6027

Web Intelligence in Information Retrieval

Kevin Curran, Cliona Murphy and Stephen Annesley
Internet Technologies Research Group
University of Ulster, Magee Campus, Northern Ireland, BT48 7JL, UK

Abstract: Web intelligence is a fascinating area in the very early stages of research and development. It combines the interaction of the human mind and artificial intelligence with networks and technology. How will the next generation Web mature? With the imminent growth of web intelligence what expectations do users have? Clearly the user will expect more from the Web than for it to merely pass raw data between people via search engines. This study attempts to define and summarise the concept of web intelligence, highlight the key elements of web intelligence and explore the topic of web information retrieval with particular focus on multimedia/information retrieval and intelligent agents.

Key words: Web intelligence, multimedia information retrieval, WWW, web information retrieval

INTRODUCTION

The web has increased the availability and accessibility of information to such a large audience that an intelligent system is required to construct a meaningful reply to a query for information. The field of study that is web intelligence involves a combination of artificial intelligence (AI) and information technology (IT) to produce an intelligent system. Web intelligence investigates the important roles that these two components have to play on the world wide web while being concerned with the practical impact they will have on the new and upcoming generation of Web empowered products, systems, services and activities[1]. The study of web intelligence draws from a range of diverse disciplines such as mathematics, linguistics, psychology and information technology[2].

The web intelligence consortium (WIC) is an international non-profit organization dedicated to the promotion of worldwide scientific research and industrial development in the era of Web and agent intelligence[3]. The web intelligence consortium identifies 9 key topics in the area of web intelligence. One of those topics is web information retrieval (Fig. 1). The 9 key topics are further divided into a total of 75 subsections. Multimedia retrieval is one of the subsections in the web information retrieval category (Fig. 2). The following definitions are an attempt to provide succinct summaries of the predominant subjects discussed in this study.

Web intelligence: is a new direction for scientific research and development that explores the fundamental roles as well as practical impacts of artificial intelligence

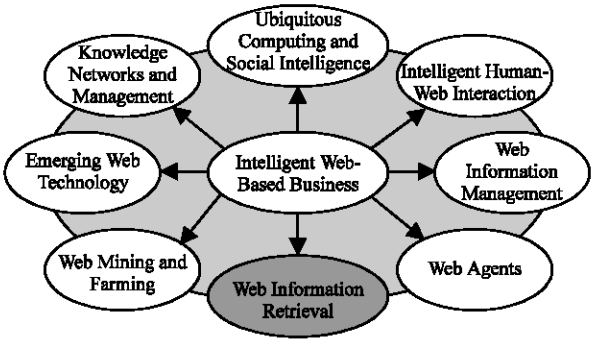

Fig. 1: Web intelligence

(AI) and advanced information technology (IT) on the next generation of web-empowered products, systems, services and activities. Goetzel describes the web of today as having "an infantile mind" and believes that over the next couple of decades we will see its growth and maturity into a fully fledged, largely autonomous, globally distributed intelligent system[4].

Artificial intelligence (AI): is concerned with the design of intelligent computer programs, which simulate different aspects of intelligent human behaviour. In particular, the focus has been on representing knowledge structures that are utilized in human problem solving[5]. In other words, AI is the simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (including the rules to reach approximate or definite conclusions and self correction). Particular applications of AI include expert systems, speech recognition and machine vision.

Corresponding Author: Kevin Curran, Internet Technologies Research Group, University of Ulster, Magee Campus, Northern Ireland, BT48 7JL, UK Email: kj.curran@ulster.ac.uk

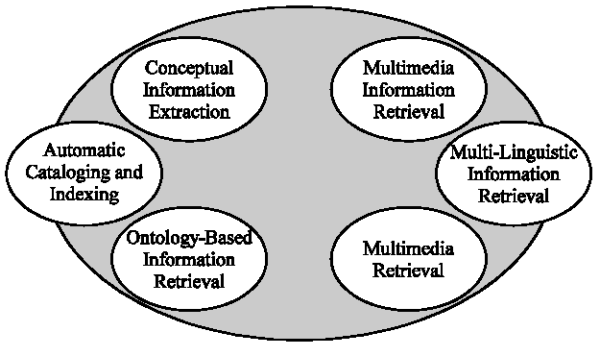

Fig. 2: Web information retrieval

Web information retrieval: Comprises conceptual information extraction, automatic cataloging and indexing, ontology-based information retrieval, multi modal information retrieval, multi-linguistic information retrieval and multimedia retrieval. This first part of the study focuses on the final subject in this list-multimedia retrieval.

Intelligent multimedia information retrieval: is a multi disciplinary area that lies at the intersection of artificial intelligence, information retrieval, human computer interaction and multimedia computing[6]. It goes beyond traditional hypertext or hypermedia environments to provide content based indexing of multiple media (e.g., text, audio, imagery, video) and management of the interaction with these materials.

MULTIMEDIA INFORMATION RETRIEVAL

With the popularity of multimedia technology, contents of the world wide web have been a lot more versatile than a few years ago. However, although more information is available on the web, the efficient and effective retrieval and management of these web documents are still very challenging research issues[7]. Intelligent multimedia information retrieval involves much more than retrieving free text, it involves systems that enable users to create, process, (e.g., index, profile) summarise, present (e.g., visualise, customise), interact with (e.g., query, browse, navigate) and organise information within and across heterogeneous media such as text, speech, non-speech audio, graphics, imagery, animations and video[6].

With the rapid development of Internet technology, the number of Internet users and the amount of multimedia information on the Internet is ever increasing. Recently, the web sites such as e-business sites and shopping mall sites deal with lots of image information. To find a specific image from these image sources, usually image database engines or web search engines are used. But, the feature based retrieval capabilities of these systems are quite limited, especially for the web images[8].

When navigating the web, with such a vast collection of linked multimedia documents, users can easily get lost in its depths. Multimedia retrieval also gives users problems in finding appropriate resources and extracting information from within multimedia documents. Text and relational databases can be searched on content and indexing terms. However to find information in images, video and speech the user is dependent on the extent of the semantic description of the resource assigned by the database indexer. We need to identify users search methods to develop the technology they will use. And of course the Web needs to become much smarter if it is to optimize its own performance, as well as, package knowledge to answer our ever-increasing questions. Some users know what they are looking for and try to satisfy their needs by following appropriate links. These users may or may not find something of interest, but may easily miss other, more relevant documents far from their current browsing paths.

There exists a great demand for retrieval and management tools for visual data, since visual information is a more capable medium of conveying ideas and is more closely related to human perception of the real world. However, image contents are more complicated to retrieve than say textual data stored in traditional databases. Image retrieval techniques should provide support for user queries in an effective and efficient way, just as conventional information retrieval does for textual retrieval.

MULTIMEDIA RETRIEVAL SYSTEMS

Many image retrieval systems have been developed; such as QBIC[9], VisualSEEK[10] and Photobook[11]. For instance, MultiMediaMiner[12] is a prototype of a data mining system for mining high-level multimedia information and knowledge from large multimedia databases. Some systems rely on key word only retrievals and others support image content-based retrievals. In the latter approach, they support image retrievals based on the image feature information, such as average colors, color histograms, texture patterns and shape objects. However, most of them are developed for image database applications. Multimedia applications, such as video conferences or web collaboration, bundle several means of communication (language, text, image etc.). Speech recognition, or speech-to-text, involves capturing and digitizing the sound waves, converting them to basic language units or phonemes, constructing words from

phonemes and contextually analyzing the words to ensure correct spelling for words that sound alike (such as write and right). The multimedia revolution brings a challenge for us to create an intelligent web capable of interacting and even understanding the user. This study only skims the surface of the topic of web intelligence however as it is still in the early stages of research; we all have a part to play in its definition.

INTELLIGENT AGENTS

Another aspect of web intelligence concerns the study and application of Web agents and Intelligent agents. An agent on the web can be described as a program that assembles information or performs some other service without your immediate presence and on some regular schedule. Usually, an agent program, using parameters provided by the user will search all or some part of the Internet, gather information you're interested in and present it to you on a predefined periodic basis[13].

One of the more important roles of information retrieval agents is that of searching and filtering information from distributed web sources. Thus, understanding and developing the correct information foraging behavior for information retrieval is a challenge. It is important to understand how people search for information[14]. Liu[14] presents a foraging agent model which takes into account web topology, information distribution and agent interest profile. They discovered that it is the unique distribution of agent interest that leads to the regularities in agent surfing behavior i.e., a power law distribution of agent surfing depth. The power lay of link click frequency is largely due to agent purposeful surfing behavior demonstrating that web regularities are interrelated. They also categorize foraging agents according to their interests and familiarities: random, rational and recurrent agents. They discovered that the regularities of agent surfing depth on pages and domains still remain the same, while a power lay of clicks frequency distribution will disappear as we move from recurrent to random agents. This result shows that the order existing in link click frequency comes from agent's content prediction ability, that is whether or an agent can determine the next step according to its own interest and current information.

Research of web agents by the web intelligence consortium can be broken down into a number of subcategories of agents. Examples include Semantic agents, Information filtering agents and Remembrance agents, amongst others.

Intelligent web agents can use the problem solver markup language (PSML)[15] to specify their roles, settings and relationships with any other services. An intelligent Web will have the ability to process and understand natural language. It must understand and correctly judge the meaning of concepts expressed in words, such as good, best and season. Further, the intelligent web must grasp the granularities of these terms' corresponding subjects and the location of their ontology definitions[14]. In addition to the semantic knowledge that an intelligent search can extract and manipulate, intelligent Web agents will incorporate a dynamically created source of metaknowledge that deals with the relationships between concepts and the spatial or temporal constraint knowledge that planning and executing services use allowing agents to self-resolve their conflicts. To solve specific problems, intelligent web agents must be able to plan. The planning process uses goals and associated sub goals, as well as constraints[14].

SEMANTIC AGENTS

Semantic agents operate on the semantic web. The semantic web operates on the object orientated model of classes and objects each with their own properties. It is an extension of the current web in which information is given well-defined meaning[16]. A semantic agent introduces the concept of ontology. Ontology is a means of describing information. It is a set of descriptors, including the vocabulary, the semantic interconnections and some simple rules of inference and logic. Ontologies let information on the web to be precisely defined. This better enables computers, using agents, to return a more meaningful set of results to the user. Conversational agents, or chatterbots, such as Microsoft Agent[17] or Virtual Personalities Inc.[18] are basically speech-activated agents that can direct a computer generated facial animation and that includes a learning element[19]. For such a system to operate a means of understanding natural language is required. This is undertaken roughly, in three stages of analysis: syntax, semantics and pragmatics. Syntax analysis is concerned with the structure of a sentence in terms of the relative positions of words and their parts of speech. Semantic analysis examines the meaning of words and begins to build an internal representation of the meaning of the sentence. This task cannot be completed without the pragmatics, that is, knowledge about the domain of discussion. An understanding of the pragmatics is needed to resolve uncertainty and fill in assumed knowledge about the domain. They can be used as front ends for database products allowing the user to provide a query in a more natural context although their use is limited.

INFORMATION FILTERING AGENTS

Current research into agents is also being undertaken in the field of e-mail filtering and automatic handling agents. These include such innovations as the approach being taken by A-Life BotMail[20]. They have developed a product that allows senders to deliver an interactive and intelligent bot that can converse with the recipient of the e-mail and intelligently present the contents of the message instead of sending a plain document. Snoop[21] is another type of e-mail agent that can automatically inspect and evaluate your incoming mail messages and based on what it finds it can take certain actions such as generate an auto response to an incoming e-mail. It can also forward e-mails, control your PC by e-mail by launching certain applications in response to an e-mail and parse messages and log information to text files for storage. Other examples of web agents include 'Copernic Agent'[22] for information retrieval on the web, 'E-mailrobot'[23], an e-mail manager and automater, allowing you to process, route, track and manage messages and 'NewsRover'[24], a tool for extracting information from usenet newsgroups automatically. The large number of information sources on the Internet present users with the hard task of gathering relevant and useful information from a query. Such a task is too difficult to solve without some form of high-level filtering of information. Intelligent agents can aid in this area, passing on to the user only those items that they are interested in. One solution to this problem is to integrate different artificial intelligence technologies such as scheduling, planning, text processing and interpretation problem solving, into a single information gathering agent, called BIG (resource-bounded information gathering), that can take the role of the human information gatherer[25]. Agents have evolved to the point that complex web agents now exist that can learn their user's preferences and actively seek out Web pages that could be of interest to them To provide personal assistance, an agent needs information about the user's interests and needs[26]. Agents can suggest information sources and products to users based on learning from examples of their likes and dislikes. Such an agent is termed a 'Recommender Agent'. Two modes of operation exist for recommender agents. Most existing recommender systems use social (collaborative) filtering methods that base recommendations on other users' preferences from web sites. By contrast, content-based methods use information about an item itself to make suggestions. This approach has the advantage of being able to recommended previously unrated items to users with unique interests and to provide explanations for its recommendations.

REMEMBRANCE AGENTS

A remembrance agent is a program that aids human memory by displaying a list of documents that might be relevant to work the user is doing. Unlike most information retrieval systems, the agent runs continuously without any user intervention. Its unobtrusive interface allows a user to follow up or ignore the agents suggestions as desired. The front end of the agent program continuously watches what the user types and reads. It then sends this information to the back end. The back end finds old e-mail, notes files and on-line documents that it thinks relevant to the user's context. This information is then displayed by the front end, in such a way as not to distract the user from their current task[27]. Agents called 'Shop Bots' can aid the user in their purchase of an item of interest over the Internet. They compare prices from a number of online stores but currently are not very comprehensive, except in the computing Industry. PriceSCAN[28] offer a service allowing the user to compare the price of any tem from a number of different online stores before making a purchase. Shopping agents can compare prices on a number of different items or they can be specific to the field that they search in. Dealpilot[29] searches over 20 online bookstores for the lowest price and then returns the results to the user.

PROFILING AGENTS

Profiling agents are used to build dynamic sites with information and recommendations tailored to match the individual taste of each visitor. The main purpose of the agent software is to build customer loyalty and profitable one-to-one relationships. Learn Sesame[30] learns about users automatically from their browsing behaviour, adapting to changes in the users interests over time. Agent programs can allow the user to have data sent automatically or pushed to their computer at regular intervals, such as every hour, or when triggered by an event, such as when a web page is updated. This is accomplished using what is known as push technology. Desktop News[31] keeps the user informed by delivering a continuous stream of news and information from chosen Web sites direct to their desktop in a compact ticker toolbar. Push technology is an alternative to the way the World Wide Web currently operates in that information is presented to the user without their intervention, where ordinarily the user goes online to search for information.

NAVIGATION AGENTS

Navigation agents are used to navigate through external and internal networks, remember short cuts,

pre-load caching information and automatically bookmark interesting sites. IBM's web browser intelligence[32] pronounced Webby is an example. In addition, agents can help in the development and maintenance of a web site. CheckWeb[33] scans user generated HTML pages and explore all the links for errors. When finished the program generate a log file with all errors it has found. IseekTraffic[34] submits a site to over 157,000 search engines, directories and links pages in its database. The software will submit all URL's to all of the top Search Engines and Directories.

The categories of agents discussed up to now have been engineered to operate in one particular field or mode of information retrieval. Agents can also undertake a variety of features, not being restricted to just one function. Agents such as Ultra Hal[35] can cover a range of functions such as remind you of important dates, start programs on your behalf, browse the Internet and answer e-mails.

PUTTING IT ALL TOGETHER

Agents with the desired capabilities for navigating the 'Wisdom web' are beginning to move beyond research exercises and may be in common use in the near future. However, there are still a number of limiting factors that must be overcome. Building autonomy into web agents is not the limiting factor. The key limiting factor is the difficulty of building and maintaining ontologies for web use[36]. The most basic need in interacting with an agent is a language in which to communicate. An agent that is truly useful must have a lot of knowledge about the problem being solved. If the travel agent does not know about geography (Where are the Maldives?), transportation (What cruise ships sail there?), economics (Can I afford the trip?) then we cannot easily communicate our needs[36]. Building ontologies is a daunting task, especially as detailed technical knowledge is needed to provide truly useful searches[37] and these ontologies must be brought to the web in a machine readable form. Major efforts are underway to overcome these problems and to develop new tools for creating ontologies and/or bringing them to the web[38]. There is reason to believe that as the current set of web tools such as SHOE[39] become more capable the bottleneck in developing ontologies will be overcome.

Improvement is also being seen in the effort to make agents more capable such as the market forces which are now driving online journals to explore the greater use of agent-based systems. Current search engines, using keyword based techniques, are inadequate for providing the detailed sort of searches needed by the scientific community. XML is being used to organize scientific material, making it easier for web agents to find key aspects of scientific documents. XML also make it easier for agents to become 'capable' as they can more clearly identify what payment is required, what information is needed for downloads, etc. The area of agent-based systems is a hot one. We have the technology to build software agents that are communicative, capable, autonomous and adaptive – the key behaviors needed to help make information retrieval more fruitful. The limiting factors in building such systems are being overcome and new approaches are emerging. IT seems that sooner rather than later, we will all be using agent-based technology[36].

CONCLUSION

The success of web intelligence will not hinge on available technology alone, but rather on the widespread acceptance of the medium to meet the needs of the user at large. Web Intelligence developers will essentially need to combine teams of people with varied perspectives to develop and design an intelligent web that will effectively address the ultimate requirements of the user. In doing so we may witness a growth and adoption of web intelligence, encompassing every area of commercial enterprise and every aspect of human endeavor, resulting in the proportional displacement of conventional methods of communication. All categories of intelligent agents discussed in this study, although diverse, have one thing in common. They are all constructed to allow the user to query the Internet and its vast array of back end databases and bring back a meaningful set of results which are relevant to the user and allow them to carry out their tasks more efficiently and effectively. Intelligent information retrieval is a small part of web intelligence that gives us the opportunity to improve the quality and effectiveness of interaction for everyone who communicates with a machine in the future.

REFERENCES

1. The Web Intelligence Consortium, http://wi-consortium.org/ (Accessed 25/5/04)
2. Zhong, N., J. Liu and Y. Yao, 2002. In Search of the Wisdom Web. IEEE Computer, 35: 27-31
3. http://wi-consortium.org/
4. Goertzel, B., 2002. The emergence of global web intelligence and how it will transform the Human Race. http://www.goertzel.org/papers/webart.html (Accessed 25/5/04)
5. Preece, J., 1998. Human Computer Interaction. Addison Wesley, UK.
6. Maybury, M., 1997. Intelligent Multimedia Information Retrieval. AAAI Press/MIT Press, London.

7. Pringle, G., L. Allison and D. Dowe, 1998. What is a Tall Poppy among Web Pages?, Proc. 7th IWWWC, Brisbane, Australia, pp: 113-117.
8. Hong, S., C. Lee and Y. Nah, 2002. An Intelligent Web Image Retrieval System. Department of Computer Engineering, Dankook University, Seoul, Korea. Technical Report Available at http://dblab.dankook.ac.kr/SPIE2.pdf, (Accessed 25/5/04)
9. http://wwwqbic.almaden.ibm.com/
10. http://www.ctr.columbia.edu/VisualSEEk/
11. http://www-white.media.mit.edu/vismod /demos/facerec/
12. Zaiane, O., J. Han and S. Chee, 1998. MultiMediaMiner: A System Prototype for MultiMedia Data Mining, Proc. of SIGMOD98, USA.
13. Browne, C., 2002. Web Agents, http://cbbrowne. com/info/agents.html (Access 25/5/04).
14. Liu, J., 2003. Web Intelligence (WI): What makes Wisdom Web?. 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, pp: 1596-1601.
15. Menasalvas, E., J. Segovia and P. Szczepaniak, 2003. Advances in Web Intelligence: First International Atlantic Web Intelligence Conference AWIC 2003, Springer-Verlag Heidelberg Madrid, Spain, May 5-6.
16. Hendler, J., T. Berners-Lee and E. Miller, 2002. Integrating Applications on the Semantic Web. J. Institute Elec. Eng. Japan, 122: 676-680.
17. Microsoft Agents, 2004. Online at http://www. microsoft.com/msagent/ (Accessed 25/5/04).
18. http://www.verbots.com/ Virtual Personalities Inc. (Accessed 25/5/04).
19. Sammut, C., 2000. Conversational agents. University of New South Wales, Australia. Report online at http://www.cse.unsw.edu.au/~claude/projects/nlp. html
20. http://www.artificial-life.com/v5/website.php
21. http://www.smalleranimals.com/snoop.htm
22. http://www.copernic.com/index.html
23. http://www.gfisoftware.com/
24. http://www.newsrover.com
25. Lesser, V., B. Horling, F. Klassner, A. Raja, T. Wagner and S. Zhang, 2000. BIG: An Agent for Resource-Bounded Information Gathering and Decision Making. Artificial Intelligence J., Special Issue on Internet Information, 118: 197-244.
26. Payne, T. and P. Edwards, 1995. Learning Mechanisms for Information Filtering Agents. Proceedings of the UK Intelligent Agents Workshop, SGES Publications, Oxford, UK, pp: 163-183.
27. Bradley J. R. and T. Starner, 1995. Remembrance Agent. A Continuously Running Automated Information Retrieval System. MIT Media Lab, Cambridge, MA. First International Conference on The Practical Application of Intelligent Agents and Multi Agent Technology (PAAM '96), pp: 487-495.
28. http://www.pricescan.com/
29. http://www.dealpilot.com/
30. http://www.aminda.com/mazzu/ls.htm
31. http://www.desktopnews.com/
32. IBM, 2004. Web Browser Intelligence-Agent Software, Online at http:/lwww, raleigh. ibm. com/wbi/wbisoft.htm. (Accessed 25/5/04)
33. http://www.algonet.se/~hubbabub/how-to/ checkweben.html
34. http://www.botspot.com/Intelligent_Agent/1986.html
35. http://www.agentland.com/
36. Hendler, J., 1999. Is There an Intelligent Agent in Your Future? www.Nature.com, March (25/5/04).
37. Heflin, J. and J. Hendler, 2000. Dynamic Ontologies on the Web. Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000), Boston, USA, pp: 120-126.
38. Ceruti, M., C. Anken, A. Lin and S. Rubin., 2000. Applications of High-Performance Knowledge-Based Technology, in proceedings of IEEE Systems Man and Cybernetics, 20: 45-53.
39. Heflin, J. and J. Hendler, 2000. Searching the Web with SHOE. AAAI-2000 Workshop on AI for Web Search, pp: 219-224.