Evolving User Profiles to Reduce Internet Information Overload

John Pagonis[1] and Mark C. Sinclair[2]

[1] Personal Area Networking, Communication Technology Group,
Symbian Ltd., 16, Harcourt Street, London W1H 1DT.
Email: john.pagonis@symbian.com

[2] Dept. of Electronic Systems Engineering, University of Essex,
Wivenhoe Park, Colchester, Essex CO4 3SQ.
Tel: 01206-872477; Fax: 01206-872900; Email: mcs@essex.ac.uk

Abstract. This paper discusses the use of Evolving Personal Agent Environments as a potential solution to the problem of information overload as experienced in habitual Web surfing. Some first experimental results on evolving user profiles using speciating hybrid GAs, the reasoning behind them and support for their potential application in mobile, wireless and location aware information devices are also presented.

1 Information Overload

In everyday life, the Internet user is faced with the ever increasing problem of information overload, whether this occurs at home, at the workplace, or as will soon be happening, everywhere [1] [2]. The overwhelming information feed that computer users face leads to anxiety, strain, inefficiency and finally results in uninformed (or misinformed) and frustrated users [3] [4]. Continuously and increasingly Internet users are confronted with laborious and difficult tasks of information filtering and/or gathering, which are inherently computer-oriented processes and need not involve human labour. Moreover, the efficiency and speed with which computer systems may retrieve and process massive amounts of data actually exacerbates the problem of ineffectiveness in information filtering without user intervention.

In our research programme, which we first presented in [5], we aim to solve the information overload problem arising from *habitual surfing*, with a focus on mobility, Wireless Information Devices (WIDs) and privacy. Therefore moving away from the model of *what is of interest to the user* to the model of *what is of interest to the user here and now*. In [5], we surveyed recent efforts at tackling the problem of information overload. Our view is that, as a means of achieving our goal, we need to research Evolving Personal Agent Environments (EPAE).

The focus in this paper is on evolving individual user profiles, which form a vital part of the EPAE research. Thus we aim to describe and explain our definition of user profiles in personal agent environments, and to support the motivation behind it. We also include an overview of our proposed research programme on EPAE and present some early experimental work on user profile evaluation and evolution.

2 Evolving Personal Agent Environments

Our means of dealing with the problem of information overload, particularly in the domain of the World Wide Web, is geared towards the use of personal and adaptive agent-based systems. The driving ideas behind such a system and an early outline of it, are presented below.

Central to the design of the suggested system are the use of evolutionary algorithms, agent technologies, collaborative filtering, privacy, agent-society trust relationships, location awareness and mobility, as well as continual adaptation [5]. Our belief is that through employing an evolutionary-algorithm-based multi-agent environment, where the 'fitness' of each individual is determined by an estimate of user satisfaction, the system should more closely adapt to the user's personality and preferences, thereby yielding increasingly better results. However, while selection of appropriate measure(s) to estimate user satisfaction will form part of our research programme, which is geared towards minimising user intervention, nevertheless it remains outside of the scope of this paper.

The system which we aim to produce will focus on transparently assisting in the filtering and presentation of information. This will be based on observation of the user's individual behaviour, enhanced through adaptation and mobility. As its major target, the proposed system will aim to constantly and continuously evolve and adapt to the locational and behavioural changes of the user. Our goal is to architect, and subsequently implement, this system which will ease the information overload problem and closely follow the individual user's interests. Our experimental work towards designing the system commenced with the evolution of basic user profiles, and is presented in the following sections.

3 User Profiles to Reduce Information Overload

We envisage that user profiling in the near future will be more and more coupled with the actual location where the user encounters information overload. Moreover active user profiles and browsing behaviour will depend upon location, and also possibly on schedule. Location awareness (or context sensitivity [6]) comes as a direct result of the use of portable computers, personal data assistants, wearable and WIDs, plus other enabling technologies, like mobile IP, GPRS, GPS and others. This ability of the mobile device, and hence of the agent environment, to discover its location, will enable new ways of following the user's behaviour, habits and interests, thus adapting and evolving more closely to them. Context sensitivity and mobility will doubtless introduce continual change and development of the users' interests yielding many profiles belonging to the same user.

In our research we define the *user profile* as a set of subjects and interests that are relevant and useful at a given location and time. For example, this could mean that a different view[1] onto the Web might be presented to the user, according to different profiles, throughout the day, *e.g.* when in the office, when at home, when on the bus. All the user satisfaction feedback, behavioural and locational changes could then be recorded. Then, when evolution next occurs, it will include all these different individual profiles. In addition, statistical behaviour could be analysed to anticipate the next locational/calendar profile change, in order to be in a position to present the user with the desired and relevant information digest. To probe further, the user could confirm creation of each different profile to allow the agent system to 'legitimately' evolve them. Though the system, based on statistical inferences, could simply make the user realise such profile changes by asking the right questions: "I see that you are on the bus quite often on Monday mornings, do you feel like reading the morning news then?"

4 Evolving User Profiles

We anticipate our approach to representing user profiles in the computer domain will be based on speciating genetic algorithms (GAs) according to location and time. In this approach, a user may have a number of domains of interest (or interest pools) which may or may not be relevant and/or useful across time and space.

Hybrid GAs were chosen to ensure that the properties of the algorithms developed would be well matched to the problem domain, and also because recent work by Moukas [7] [8] [9], Chen *et al.*[10] [11], Chen & Sycara [12] and others has shown that GAs can offer promising solutions. Moukas has demonstrated that GAs acting as filters (part of a wider ecosystem) can be very responsive in adapting to changing user behaviour [7] [8] [9]. Moreover Chen's research suggests that filtering using GAs is effective and can actually perform better than conventional techniques if the search criteria or the interest domain is vague or with no well defined starting point [10] [11].

Hence we can employ hybrid GAs with individuals consisting of variable vectors of (initially) words and their weights, within a given Web page. These individuals will merely describe subjects of interest or simply WWW content. They can then be separated into sets according to their significance in the user's everyday life, thus establishing the context within which they are valid. This context, which characterises their species, can be defined in terms of location, time and possibly date.

[1] Termed 'personal portal' in [5].

Traditionally a user's profile was described by modelling his/her interests regardless of context, whereas our proposal augments this with information about space and time [9] [13] [10] [12] [14]. Effectively then, the representation of a user profile becomes the collection of individuals that belong, using GA terminology, to the same species, although only the fittest of these will influence the eventual views presented to the user. Moreover, because profiles are time and location sensitive, a user can have many different profiles.

Following the assumption that a user will be interested, according to context, in a finite number of 'interests', we seek to discover whether the separation of the species can speed and enhance the selection and presentation of relevant information. In this approach, utilising contextual information, the user profile becomes a collection of the interests (species) that satisfy the active context's requirements. Thus when the system is called to filter an offering of documents, rather than just employing the fittest 'filters' to decide on those most relevant for the user to read, instead not only the most fit, but also the most *relevant* filters will be used to pick the documents. Therefore, as has been said, parting from the model of what is of interest to the user towards the model of what is of interest to the user *here and now*.

4.1 The Experiment

For simplicity, in this experiment we considered the case where a user has only one active profile at any time, of a set of two, which we conveniently divided between office hours and the user being at home. Moreover each of these two profiles encapsulates a single interest (thus being a simple profile). We first evolved the profiles and they were then evaluated by the user.

To seed the initial population, a number of recent documents was selected which had been stored as links from the user's WWW browser's bookmarks under the same bookmark category, relevant to the active profile at the given time and location (*i.e.* home). Each such document would be used to generate a single individual.

Then we randomly picked from the same category another set of documents (the training documents) which would be used to evolve the population created from the former set. Each generation, the fitness of each individual from the evolving population was evaluated against a randomly picked document from this latter set.

GA Features The genotype of the individuals resulted from representing the documents as weighted keyword vectors (which we aim to augment with the contextual information as the system is developed). The Java framework which we used to build the GA application employs a design from previous work by one of the authors [15] [16], and is intended to be generic enough to accommodate any encoding that may be needed in our future work. Based on the current encoding, the fitness function was taken to be the cosine of the angle between the multidimensional vectors representing the training and evolving individuals. This technique has been successfully used elsewhere [9]. The crossover operator used was a random two-point crossover which exchanged a part of one genotype of random length with a part of the other genotype also of random length. Moreover, the mutation operator randomly changes the value of a gene's allele in the genotype. Both of these operators were assigned a low probability of 20% and 5%, respectively, and the remainder was allocated to a reproduction operator which would clone the fittest individuals.

Results and Observations The decision to assign a relatively small probability to the crossover operator was based on the observation that a high crossover probability eventually seemed to harm the population after some generations. This appeared to be the case because the operator could potentially change the size and structure of an individual considerably and since that individual represented a document filter, the filter itself could be changed to match very different documents.

We also experimented with having adaptive operator probabilities, where the probability of selecting an operator would increase with greater improvements in population fitness. That technique proved to be very unfruitful, in this case, as the population regressed rapidly, apparently due to the effects of the increased mutation. On the contrary, by keeping the operator probabilities constant and low, better results were obtained. With such operator probabilities, our experiments showed that each profile's population (representing in effect niche filters) could be rapidly evolved to a fitter population in a matter of few generations.

Moreover that population could reliably be used to select documents of interest out of a wider set of documents. To verify this, the user selected from the Web a small set of documents of interest, in alignment with his active profile. These documents were injected into a larger set of randomly selected documents, not matching that profile. The best filters were then called to sort these documents in order, ranking the most 'interesting' documents at the top of the list.

Problems and Future Directions The problem in validating and measuring the success of such experiments is that the user is involved in deciding the validity and effectiveness of the information filters, represented by the fitter individuals. The challenge is to make the objective fitness assigned to the profile population match, what is in effect, the subjective fitness given by the user. It is also very difficult to statistically build an experiment large enough as to guarantee impartiality and non-bias by the subjects participating in the experiment. Therefore we will be looking into this difficult area in the future in order to establish an appropriate methodology for grading such experiments.

On the technical side, we are looking for further improvements in many areas of the GA, such as the operators and fitness function. In particular the crossover operator, which seemed to damage the population if assigned high probability, could be re-implemented to take account of individuals' size. That change could possibly allow for more crossover operations to take place, since it would alter the embedded structure and size of the individuals less.

Moreover we wish to experiment with other fitness functions, and possibly improve the one used by normalising the alleles of the 'genes' in the two chromosomes according to their relative values. Such a fitness function would then allow two documents of different sizes to be compared for similarity with no loss of accuracy.

Another area of improvement is to apply better pre-processing to the documents, before they are encoded as chromosomes, so as to eliminate noise. Currently prepositions, articles, conjunctions and pronouns all make up the 'noise' in the evolved individuals and need to be removed from their genotypes.

As was said above, to judge whether a population (*i.e.* a species of a single interest) can successfully filter documents and suggest those of interest, its best individuals are used to sort documents in order of relevancy. This is a useful approach, but the inherent problem in applying it, is that a 'global' threshold above which documents are relevant, cannot be predicted. Thus if the system is to present only the most relevant documents to the user, it must determine how many to select, based on how many are really relevant, which is an unknown. Estimation of such thresholds is another topic for future work.

5 Conclusions

Our motivation behind classifying user profiles according to location and time is not only based on the realisation that information overload may arise when a user has to retrieve and filter useful information from a vast offering, but also that people may now simply have interests in too many things. Moreover, that these are not always necessarily useful or relevant to their activities.

User profiling, in the most general context (of the Web domain), involves recording and somehow modelling users' preferences in viewing and navigating through webs of hyper-linked documents. However, both content preference and navigation describe people's behaviour and interests, and hence raise privacy and confidentiality issues. We will consider these in our research, and in the future we hope to examine how they can be addressed. This requirement of privacy will be even more critical if our research on modelling users' preferences, as described, proves to be a success.

The early work presented here indicates that we can obtain rapid evolution of species(*i.e.* small populations) which provide effective niche contextual filtering. It is to be hoped that eventually the system will allow us to:

- evolve newly discovered profiles on-the-fly;

- quickly apply the niche filters/populations representing interests to documents;
- give user satisfaction with minor computational overhead;
- employ a small memory footprint for both storing and processing;
- require minimal overhead in encrypting profiles;
- effectively transport the personal profile information over low bandwidth, unreliable or Personal Area Networks (PAN)

Although our research aims may seem futuristic, it is apparent that mobile platforms like WIDs will soon be able to support environments allowing research systems like ours to be realised. Evolving user profiles may then become commonplace in a fast moving information-device-laden society, where adaptability and mobility are as ubiquitous as telephony and mobile phones are today.

References

1. Halla, B. (1998): How the PC will Disappear, IEEE Computer 31, 136–138
2. Kleinrock, L. (1999): Nomadic Computing and Smart Spaces, IEEE Internet Computing 4, 52–53
3. Ireland, P. (1999): Management Paradigm to Reduce Information Overload, Computing & Control Engineering Journal 10, 29–32
4. Bowman, C.M., Danzig, P.B., Manber, U. & Schwartz, F. (1994): Scalable Internet Resource Discovery: Research Problems and Approaches, Commun. ACM 37, 98–107
5. Pagonis, J. & Sinclair, M.C. (1999): Evolving Personal Agent Environments to Reduce Internet Information Overload: Initial Considerations. In: IEE Colloquium on Lost in the Web: Navigation on the Internet, Digest No. 1999/169, London, UK, 2/1–2/10
6. Billinghurst, M. & Starner, T. (1999): New Ways to Manage Information, Computer 32, 57–64
7. Moukas, A.G. (1996): Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem. In: Conf. on Practical Application of Intelligent Agents and Multi-Agent Technology, London, UK
8. Moukas, A. (1997): Amalthaea: Information Discovery and Filtering Using a Multiagent Evolving Ecosystem, Applied Artificial Intelligence 11, 437–457
9. Moukas, A. & Maes, P. (1998): Amalthaea: An Evolving Multi-agent Information Filtering and Discovery System for the WWW, Autonomous Agents and Multi-Agent Systems 1, 59–88
10. Chen, H.C., Chung, Y.M., Ramsey, M. & Yang, C.C. (1998): An Intelligent Personal Spider (Agent) for Dynamic Internet/Intranet Searching, Decision Support Systems 23, 41–58
11. Chen, H.C., Houston, A.L., Sewell, R.R. & Schatz, B.R. (1998): Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques, Journal of the American Society for Information Science 49, 582–603
12. Chen, L. & Sycara, K. (1998): WebMate: A Personal Agent for Browsing and Searching. In: 2nd Intl. Conf. on Autonomous Agents and Multiagent Systems, Minneapolis, USA, 132–139

13. WebMate, http://www.cs.cmu.edu/~softagents/webmate.html
14. Autonomy, Technology White Paper, http://www.autonomy.com/tech/wp.html
15. Sinclair, M.C. (1998) Operator-probability Adaptation in a Genetic-algorithm/Heuristic Hybrid for Optical Network Wavelength Allocation. In: IEEE Intl. Conf. on Evolutionary Computation (ICEC'98), Anchorage, Alaska, USA, 840–845
16. Sinclair, M.C. (1998): NOMaD: An Optical Network Optimisation, Modelling and Design Toolset. In: IEE Colloquium on Multiwavelength Optical Networks: Devices, Systems and Network Implementations, Digest No. 1998/296, London, 6/1–6/6