User focus management in agent worlds

W.Pasman, 11/4/3

Introduction

Emerging distributed agent systems, smart buildings, grid computing and many other architectures leave the user lost between a big heap of agents. The popular 'yellow-pages' solution to agent discovery will brake down when multiple highly-specialized agents are needed for even the simplest actions. On the other hand, a personal agent knowing exactly what the user wants and negotiating with the agents exactly to get the job done seems technically extremely difficult to realize. In this report we discuss the various in-between solutions that have been proposed in literature.

Furthermore, an agent world is a very service-oriented world, which is hard to connect to a goal-driven user. Or to put it differently, it is not clear how the user, or an agent working for him, can coordinate the appropriate agent(s) to handle the user's goals and needs.

We start with an example to illustrate our problem. Then we discuss existing work in the relevant areas.

Example: Turn on the Light

Let's start with an example, to illustrate our problem. Assume the user wants to turn on the light of the painting illuminating a picture in a museum. We look at three scenarios, one where a personal agent (**PA**) is helping the user to find the service, one where there is no help from the system, and one where agents are associated with world objects and have their own interface available.

In the PA-knows-it-all scenario, the user just says "illuminate this painting brighter" to his personal agent. The personal agent then would find the relevant light-agent and send it the right command to the agent, in a precisely defined agent communication language format (**ACL**). In this extreme, the personal agent would have to know all about almost anything - for instance the typical words used in a 'light management' task, the semantics of "brighter" and "this painting", how these words translate to an ACL command, etc. The knowledge in the personal agent knowledge has to be extended if a new service (e.g., shoe repair agent) is introduced to the personal agent. In an even more advanced scenario, the PA would know that the user has bad eyesight (or wants to take a picture) and turn up the light even before the user arrives at the painting.

In the user-manages-all scenario without system support, the user would start with searching an agent directory service that lists all available agents for him. Using a search facility in the agent directory service, he then searches for

lights in the environment. Then he has to ask about the language and ontology [Uschold96] that the lamp speaks. Most lamps probably return their 'location', but using different ontologies because there are different types of lamps and each brand has defined its own ontology. So the user has to download and understand the ontology definitions in order to understand the location of the lamps and to find the right one, and to understand the mechanisms behind brightness changes. Furthermore, the user has to find out whether he has access rights to the light, and how to negotiate with the lamp about pricing and access rights. Clearly in this scenario turning on the light has become a near impossible task.
In an 'environment-manages-it' scenario, there might be a microphone attached to the picture. The user asks the painting if it can show itself brighter. However he might also try to shout at the lamp high in the ceiling. The museum might track the user, and if it has access to the user's profile it can also pre-adjust the lights to fit the user. This scenario can be made to behave similar as the PA-knows-it-all scenario or the user-manages-it-all scenario, depending on how much intelligence and coordination there is in the environment.

If we consider the "PA-knows-it-all" and the "environment-manages-it" as two extremes, there would be an intermediate solution where the PA searches and finds the services that the environment provides and assists the user in addressing these services.
Given the assistance of the PDA, it is still unsolved how users would address multiple agents simultaneously, i.e. interaction is in time-sharing mode and it should be obvious to the PDA or the environment when the user switches contexts.
In person-to-person communication the user would look to the addressed person. In a WIMP interface, the user would click on another window. How is the context-switching done in a mobile situation with only a PDA at hand?

Related work

In this section we discuss work related to our problem of connecting the user to relevant services and communicating with them.
We start with a discussion of physical location as the basis for user focus estimations. Then we discuss the use of the user's gaze and where his eyes are pointing to. We end with clues from his utterances and actions.
It is unfortunate that the keywords that seem relevant to our topic are so overloaded that mostly irrelevant reports come up when we tried to locate relevant research. We did numerous searches but did not succeed in finding good work on user focus management. This may either imply that there simply is no such work, or that we just could not find it.
In general it seems difficult to compare and evaluate complex, multimodal systems [Beringer02]. Dialog strategy differences between subjects, different hard- and software, scenarios, fuzzy user task definition (because they are

expected to solve it themselves, in cooperation with the system) etc are a few of the problems that are hard to abstract away from.

Physical location as focus

Smart or intelligent houses enable the user to access all his house appliances electronically, such as the lights, heating, blinds, tv, video recorder, refridgerator and magnetron. Most smart house systems use the user's location to provide him a reduced set of available appliances. Speed [Speed97] provides the user a list of available appliances on a tablet PC, so that he can pick the targeted appliance by hand. Although her user study showed that this works quite convenient, it will become difficult to find the right appliance if the number of appliances gets too large. Microsoft's EasyLiving system [Brumitt00] uses similar techniques. They have a wireless mouse automatically grabbing the pointer on the nearest screen. Audio and video devices are automatically selected using distances between user and appliances, provided by the system's geometry services. All devices have their own XML page which is displayed if the user wants to communicate with the device. Displays and speakers can be changed automatically if the user moves around. From the paper it seems that the user just has to browse through a hierarchy of available devices to select an appliance, for instance the 'room control' to control room settings. Geometry can be used to select lights (instead of the actual light name eg 'light37'), but it seems that actually the list of lights is only restricted with the current user location, and that the user still has to pick the right one from the filtered list.

More focused on the office environment, AT&T Developed the "active badge system" [Want92], which tracks people in a building. It can be used to forward telephone calls, to inform coffee automata about user preferences, for billing, etc. Ward and Addlesee [Ward01], also at AT&T developed a similar but cheaper system where the user's computer is automatically forwarded to a terminal when they get close to it. In these systems it seems that the user can't directly access the service of his choice, but that he has to move there physically.

Asthana et al. [Asthana94] developed a shopping assistance system. The system tracks the shop visitor, can inform the shopper where articles are located, shows location dependent commercials and can warn for wet floors. They use voice recognition, for instance the shopper can say 'peas' in the personal device he gets at the entrance of the shop, and the system may respond with "fresh vegetables are in aisle 15. Frozen vegetables are in aisle 8. All Green Giant products are 25% off today". The system also can give background music (or news or stock prices) to the user. It is not clear to what extent this system has actually been implemented.

A lot of research is going on on the intelligent building concept, but unfortunately most reports don't report their solution to the appliance selection problem.

Gaze and voice as focus

Tracking the user's gaze is a popular means to determine the user's current 'focus'. Head tracking can be used in a straightforward way, by using head movements to replace a mouse [Medl98]. However its use is restricted, as a mouse pointer restricts the head tracking information to the WIMP interface, allowing the user only to select items (windows, icons) on the screen.

More interesting, Stiefelhagen [Stiefelhagen02] proposed to use the user's gaze to determine which kitchen or home appliance the user is addressing or which meeting participant the user is talking to. Stiefelhagen reports about 73% accuracy of the focus estimation using the user's gaze direction. Stiefelhagen reports that he has gaze tracking difficulties because the users are not always at the same position when they address a device, nor do they necessarily look at it. Maglio and Campbell [Maglio03] tracks the words that the user's gaze hits on the screen for determining current context, and uses this context to select news items that are shown on a ticker display. Medl et al. [Medl98] used gaze direction to detect which object the user was referring to when he refers to it with voice ('move **this** object'). Similar techniques have been used in the SmartKom project [Wahlster02].

Shell, Selker and Vertegaal [Shell03] automatically select the right vocabulary for voice recognition, depending on the object the user is focusing on. This assumes that the object is a clear representative of the task that can be done. They also use this concept the other way around, with virtual eyeballs that focus on the user if someone is trying to reach the user.

There are some reports on estimating the user's intentions from his gaze [Goldberg95, Salvucci99]. Actually the term 'user intentions' seems to be overstretched in these cases. For instance Salvucci [Salvucci99] has the user look at keys on a picture of a keyboard shown on the screen. He uses the term "estimation of user intention" for his probabilistic procedure that determines which word in a list of words is closest to the letters the user looked at. I'm not aware of systems that really use the user's gaze direction on a plan and goal level (discussed below).

For some applications head tracking (that is, the position and orientation of the head alone) is sufficient, while for others more accurate gaze tracking (where the eyes are actually looking) is required. For Stiefelhagen's meeting application rough, face-based gaze direction estimations were sufficient in most cases to determine to who a user was looking. Such face-only based tracking results in typical errors in the order of a few degrees from the exact direction where the eyes are focused. In other cases, such as in the system of Maglio and Campbell, the user's pupil is tracked to determine exactly where he is looking.

Environment influence on attention

The environment will influence and maybe even steer the user's focal attention. A huge amount of research has been done on where the user focuses, how he traverses through the visible objects, how contrast and luminance influence this, etc. To give a few examples, it is known that audio

often draws more attention than video and that it is also very quick in drawing attention. Stiefelhagen [Stiefelhagen02] showed that detecting which people are speaking in a meeting can be just as effective for determining the user's focus as the user's gaze, and combining the user's gaze and sound gives significant improvement of the focus estimation over voice or gaze alone. A large amount of research has been done on where the user focuses, how he traverses through the visible objects, how contrast and luminance influence this, what the effect is of symmetry, etc. Space here is too restricted to even start giving an overview of this vast area.
Clearly, attention is not a one-way road from user intention to gaze direction, the environment is very well able to at least direct the user's gaze direction. Probably the environment influences the user's goals and intentions as well. If the user's eye focus is not depending on the user's intentions alone, but also steered by the environment, should we correct for these environmental issues to get the real user intentions? Or are the user intentions just not a stand-alone thing, and are they partially determined by the environment as well?

Affordances
A question related to the user's gaze is how the user would be able find out abilities of his environment (**affordances**) in order to constuct plans to fulfill his goals and needs. In our example, how can he know that he can talk to the painting to get it brighter. A microphone visible at the corner of the picture would give a clue, but as already said the microphone does not have to be visible or even at the painting. It could also be his mobile phone forwarding his voice to the painting-agent.
The usual approach seems to use the physical objects as an intermediary to the interfacing agent. Nearby objects are already used to guide the interface, and the next logical step is to talk to them directly. However, this approach using physical intermediaries seems not to solve too much. First, a corresponding physical object may be hidden, for example a light in the ceiling, or computer safely behind a steel door. That is, a physical object may be an intermediary for not so obvious other things, in our example the painting may intermediate to the light. A Second, related problem is that it may be not very clear where the natural place for an agent is. For instance many people think that a computer is its display and have difficulties understanding that you can have a computer without display or with two displays. Would you talk to a plant, the household robot or the tap if you intend to water the plant? Third, many agents are exactly to get rid of the physical counterpart (agenda, notepad, electronic files, the PA), and some don't even have a physical counterpart (the video recommendation agent, the voice-to-text agent). Fourth, one of the driving forces behind the agent technology is the potential to have agents hopping between machines, which further lowers the coupling of agents with a physical object.
Part of this problem is known as the 'disappearing computer' problem. However, the real problem with the disappearing computer is sometimes

(and not very accurately) called 'mental disappearance' [Streitz01]: the user may never know that a service is available if nothing suggests its existance. Even if a computer is visible, the user may not know what that makes possible.
The next section will discuss these problems in more detail.

Non-physical or distant objects

As we already concluded in the previous section, there are a few reasons to consider situations where the user is focusing on an object that is not directly visible. The user may be trying to address or objects that are not in the vicinity. Furthermore, non-physical objects may be addressed by the user. Finally, it can be argued that the interaction with physical objects is only the final step in achieving a goal. Users may be helped better if higher-level goals are detected early, next actions may be predicted better, errors in the user's plan may be recognised, etc.

Distant objects

It is clear that the methods from the previous section will fail in such a case. What can be done is to use objects in the current user's environment as a substitute for the objects that he is talking about.
People use language to transform everyday objects into stand-ins for more complex, abstract, or less tangible objects, thereby extending or 'augmenting' their context and the range of potential meanings. For instance in a military setting a coffee mug may represent a water tower, and a pen a major road [McGee01]. In the army post-it notes as a placeholder for a tank or site are very popular instead of computers, because they are robust, high resolution, malleable and can be put and kept on the map. In the Agora project [Streitz01] any kind of information or meaning can be attached real objects. However, the user will have to remember what the placeholder was for, and the computer has to uniquely identify the object in order to bring in the attached data.

Services

Another class of invisible objects are services, such as payment, language translation, money conversion or travel information. Traditionally such services were 'attached' to a human clerk, but many of these services have become computer programs. As we discussed in the introduction, users may quickly get lost in such services.
The border between such services and the manipulation of physical objects is not very clear-cut. For instance the settings of a very physical computer display is already a 'software' service.
Users have to learn the task- or service-related vocabulary or ontology anyway, being it typical object shapes, icons, menu items, words, gestures,

actions or artefacts that listen to voice. Learning the ontology has always been required when someone starts using a new machine or mechanism, but the learning of affordances has become especially urgent in computers, with invisible, cryptic text commands and fastly changing WIMP computer interfaces. Switching to a natural language or gesture interface may make these problems even worse, because these commands have a high potential for ambiguity, resulting in unpredictable computer behaviour.
One solution is to put a human secretary between the services and the human. The secretary mediates between the potential fuzzy user requests, and 'solves' the affordance problem by being a kind of universal mediator that the user can ask about anything. KPNs Eileen is an example of such a service [YPCA02]. Numerous attempts have been done to make a similar but electronic secretary ([Conita01], [Executive02]). Usually these are much less flexible, have very simple menu-like dialog schemes, a very limited user preference mechanism and no notion of user intentions. The original Cactus proposal was in this line of thinking [Pasman02c], attempting to add user preferences, learning mechanisms, context sensitivity and user intentions to an electronic secretary. However this approach has some technical problems, as the secretary needs an enormous amount of knowledge in order to properly track the user's focus, translate user request in proper messages for agents, to synchronise various agents, etc.
Alternatively, questions about user's focus, interface issues, negotiation protocols etc could be distributed over the services themselves. Although this seems a feasible approach, we are not aware of existing research using such an approach to user focus.

Goal and plan recognition
Having the system work on a goal and plan level is convenient. At that level, user actions can be predicted, alternative plans can be suggested, feasibility of plans and goals can be checked, user misconceptions can be detected, etc. It also allows to integrate context information, user preferences etc in a more straightforward way. In a goal and plan mechanism, user focus can be seen as the current active path through the goal and plan graph, where the user's current action is just the surface detail.
In such a view on user focus, there are two ways to 'sense' the focus (the user's plans and goals):

1. Estimate the user's goal from his actions.
2. Bring the dialogue with the user closer to the user's plan and goal level, for instance by having the user specify the goal and target instead of providing him with a bunch of actions to pick from.

The first is the common approach. Scripts or task graphs are needed, encoding the relation the system expects between the user's actions and his more general goals and plans. The (history of) user's actions are then correlated

with these expectations, to find the most likely plan(s) and goal(s). Once found, the advantages of goal recognition as mentioned above can be exploited. Often, the user goal and plan can not uniquely identified, resulting in reduced effectiveness of the potential advantages. Collagen [Rich00] is a typical example using this approach. Alternatively, it is possible to try to fit the user's utterance in all available parsing frames, and then pick that one that best catches the utterance [Blaylock01]. More advanced approaches use semantic networks and marker passing to understand the relations between the user's utterances [Norvig87] or make a logical derivation of how the user's actions might fit in the user's plans and goals [Wilensky02]. We discussed these in more detail in our previous report [Pasman02].
Systems using the second approach negotiate directly about plans and goals. For example the TRIPS interactive planning system [Blaylock02] assumes that planning is a complex task that the computer can not do on its own, but that the computer can help him working out the details of a plan to find the consequences, which may give rise to alterations of the plan. In the TRIPS scenarios, the user proposes some action to be taken at a relatively high level, for instance he can say "let's use the helicopter to pick up the people from X". The system then tries to fit the time schedule, resources, and other plans to fit this, and indicates what the effects of this plan are. In such a system it becomes unclear what the user focus is in the sense of goal and plan graph, as the plan graph is partially worked out by the system and does not require active user focusing anymore. However the user will still check the outcome of the plan graph in terms of resource use (time, number of helicopters and other transport means available, amount of fuel).
It is not clear whether the user is really helped with getting him into a goal- and plan level negotiation instead of direct manipulation. The questions about efficiency, mental load etc probably apply here too, and indirect manipulation very well might give a higher mental load as compared to direct manipulation.

References

[Asthana94] Asthana, A. Cravatts,M., & Krzyzanouski, P. (1994). An indoor wireless system for personalized shopping assistance. In: L.-F. Cabrera & M. Sattyanarayanan, Workshop on Mobile Computing Systems and Applications (IEEE Computer Society Press, December), 69–74. Available Internet: ucsc.edu/pub/wmc94/asthana.ps.

[Beringer02] Beringer, N., Kartal, U., Louka, K., Schiel, F., & Türk, U. (2002). PROMISE - A Procedure for Multimodal Interactive System Evaluation. Technical Report 23, SmartKom Project, Ludwig-Maximilians-Universität, München, Germany. Available Internet: http://www.smartkom.org/reports/Report-NR-23.pdf.

[Blaylock01] Blaylock, N. (2001). Retroactive recognition of interleaved plans for natural language dialogue. Technical Report 761, University of Rochester, Department of Computer Science, December. Available Internet: http://www.cs.rochester.edu/u/blaylock/Pubs.

[Blaylock02] Blaylock, N. (2002). Managing Communicative Intentions in Dialogue Using a Collaborative Problem-Solving Model. Technical Report 774, University of

Rochester, Computer Science Department, April. Available Internet: http://www.cs.rochester.edu/u/blaylock/Pubs/./Files/tr774.pdf.

[Brumitt00] Brumitt, B. L., Meyers, B., Krumm, J., Kern, A., Shafer, S. (2000). EasyLiving: Technologies for Intelligent Environments. Handheld and Ubiquitous Computing, 2nd Intl. Symposium, September, 12-27. Available Internet: http://research.microsoft.com/barry/research/index.htm.

[Conita01] Conita (2001). Conita's Personal Virtual Assistant. Available Internet: http://www.conita.com/pdfs/conitafactsheet.pdf.[Executive02] Executive Services International (2002). Personal Assistant Services. Available Internet: http://virtual-offices-online.com/secretary.htm.

[Goldberg95] Goldberg, J. H. , & Schryver, J. C. (1995). Eye-Gaze Determination of User Intent at the Computer Interface. In Eye Movement Research (J. M. Findlay et al., Eds.). Elsevier Science.

[Maglio03] Maglio, P. P., & Campbell, C. S. (2003). Attentive Agents. Communications of the ACM, 46 (3) (March), 47-51.

[McGee01] McGee, D. R., Pavel, M., & Cohen, P. R. (2001). Shifting Contexts in Invisible Computing Environments. CHI Workshop on Distributed and Disappearing UIs in Ubiquitous Computing. Available Internet: http://www.teco.edu/chi2001ws/proceedings.html.

[Medl98] Medl, A., Marsic, I., Andre, M., Liang, Y., Shaikh, A., Burdea, G., Wilder, J., Kulikowski, C., & Flanagan, J. (1998). Multimodal Man-Machine Interface for Mission Planning. Symp Intelligent Environments (March 23-25,Stanford University, Stanford, CA). Available Internet: http://www.caip.rutgers.edu/~medl.

[Norvig87] Norvig, P. (1987). A unified theory of inference for text understanding (Technical Report CSD-87-339). Berkeley, CA: Computer Science Division, University of California, Berkeley. Available FTP: ftp://sunsite.berkeley.edu/pub/techreps/CSD-87-339.html.

[Pasman02] Pasman, W. (2002). The Cactus UseT Architecture: Overview of relevant aspects. Internal Report, Cactus Impulse project, Delft University of Technology, 3 September. Available Internet: http://www.cg.its.tudelft.nl/~wouter/publications/publ.html.

[Rich00] Rich, C., Sidner, C. L., & Lesh, N. (2000). COLLAGEN: Applying Collaborative Discourse Theory to Human-Computer Interaction. AI Magazine, Special issue on Intelligent User Interfaces. Also available as Technical Report TR2000-38, Mitsubishi Electric Research Laboratories. Available Internet: http://www.merl.com/projects/collagen.

[Salvucci99] Salvucci, D. D. (1999). Inferring intent in eye-based interfaces: tracing eye movements with process models. Proc. Conf on Human Facotrs in Computing Systems (SIGCHI, Pittsburgh, Pennsylvania), 254-261.

[Shell03] Shell, J. S., Selker, T., & Vertegaal, R. (2003). Interacting with groups of computers. Communications of the ACM, 46 (3) (March), 40-46.

[Speed97] Speed, C. (1997). Smart House User Interaction Project Report. Available Internet: http://www.fysh.org/~perdita/Claire/Work/Projects/shui.html.

[Stiefelhagen02] Stiefelhagen, R. (2002). Tracking and modeling Focus of Attention in Meetings. Ph.D. Thesis, Universität Karlsruhe, Germany. Available Internet: http://isl.ira.uka.de/~stiefel/thesis.html.

[Streitz01] Streitz, N. A. (2001). Mental vs. Physical Disappearance: The Challenge of Interacting with Disappearing Computers. CHI2001 workshop on Distributed and Disappearing UIs in Ubiquitous Computing. Available Internet: http://www.teco.edu/chi2001ws/20_streitz.pdf.

[Uschold96] Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, Methods and Applications. The Knowledge Engineering Review, V.11, N.2, 1996. Available Internet: http://citeseer.nj.nec.com/uschold96building.html.

[Wahlster02] Wahlster, W. (2002). Multimodale Interaktion und Interface Agenten: Trends für Morgen und Übermorgen. Keynote speech, UseWare'02 (Darmstadt, June 12). Available Internet: http://smartkom.dfki.de/eng/start_en.html.

[Want92] Want, R., Hopper, A., Falcão, V., & Gibbons, J. (1992). The Active Badge Location System. ACM Transactions on Information Systems, 10 (1), 91-102. Available Internet: http://www.cs.colorado.edu/~rhan/CSCI_7143_002_Fall_2001/Papers/Want92_ActiveBadge.pdf.

[Ward01] Ward, A., & Addlesee, M. (2001). RFID Tagging People. Research project at AT&T Cambridge. Available Internet: http://www.uk.research.att.com/location/rfidtagging.html.

[Brumitt00] Brumitt, B. L., Meyers, B., Krumm, J., Kern, A., Shafer, S. (2000). EasyLiving: Technologies for Intelligent Environments. Handheld and Ubiquitous Computing, 2nd Intl. Symposium, September, 12-27. Available Internet: http://research.microsoft.com/barry/research/index.htm.

[Wilensky02] Wilensky, R. (2002). An AI Approach to NLP. Lecture notes for CS288, Computer Science Sivision, University of California, Berkeley. Available Internet: http://www.cs.berkeley.edu/~wilensky.

[YPCA02] YPCA (2002). Eileen: Your Personal Call Assistant. YPCA, Leidschendam, the Netherlands. Available Internet: http://www.ypca.nl.