Self-Organising Collective Knowledge

Introduction/Overview

Aim: This project aims to develop techniques to enable the self-organisation of collective knowledge. The form this would take is an interactive knowledge base held at a web-site that develops its own coherent structure using evolutionary algorithms in response to use inputs, browsing patterns and internal measures on consistency and importance. It is the combination of evolutionary algorithms and user interaction that holds out the prospect of combining the immediacy of discussion lists with an emergent and accessible structure.

End-product: The end-product of the project would be a combination of algorithms, design guidelines, prototype software and simulations to enable others to design, test and tune actively self-organising knowledge bases of their own for different communities and purposes. Such a technology would enable the setting up of "active-FAQs" where the semi-automatic collection and structuring of practical knowledge can be built from user questions and answers at minimal cost. This could revolutionise the building up and sharing of knowledge across diverse virtual communities.

An example application: At the moment if one wanted to set up a database of useful information on, say, techniques of fruit farming one has a choice of either relying on interactive but chaotic channels such as a newsgroup or finding the resources to maintain a useful database of information by hand. A successful SOCK project would enable a self-organising knowledge base to be built and maintained in the following manner: (1) an archive of posts of the relevant newsgroups and mailing lists would be built up (2) this information would be given a preliminary structure using pseudo-semantic algorithms (3) the knowledge base would be extended by users contributing new questions and answers (4) the knowledge base would use the information contained in the posts and the browsing patterns of users to evolve appropriate structures (5) users seeking established information will be facilitated by the structure to find what they require.

Key research problem: Although there has been significant research in algorithms for the self-organisation of knowledge as well into conferencing and internet-based knowledge access tools, there has been very little investigation into what occurs when a group of people interact with a self-organising knowledge base. Thus the prime research problem is to discover enough about the possible algorithm-user interactions to implement an effective self-organising knowledge base.

Methodology: The project will proceed via iterations of cycles of: design; implementation; user trial with extensive data collection; simulation; and evaluation. The simulation is a very important part of the methodology, this is an agent-based social simulation (ABSS) designed to capture the interactions of the human users with the adapting knowledge-base. This simulation is then used to try out ideas for the improvement of the underlying algorithms 'off-line' to inform the design in the next cycle. In this way maximum use can be made of the user trials.

The Goal

The basic idea is to support the emergence an organised collective knowledge base, i.e. develop an information system consisting of computers, software and people and connected by communication links, that achieves more than its parts working in isolation. Rather than having each individual add his or her contribution to a knowledge store, the system will support the creation of a network of interlinked concepts that constantly adapts to the needs of its users, both individually and collectively. This network would be similar to a collective neural network or "brain", that self-organises and learns from the way it is used. Thus the aim is not merely to implement another conferencing system for users to interact with each other or another aid to browsing and searching for knowledge but to exploit the combination of user interaction and self-organising algorithms so that the collective knowledge base can evolve as an entity in itself.

Methodology

Although there are some analogical models for the emergence of knowledge due to the interaction of individuals through traditional media from the philosophy of science and elsewhere (foundationalist, evolutionary, coherentist etc.), and there is a growing body of knowledge about the properties of learning algorithms, almost nothing is know about what happens when the two are combined. For this reason this project will have to conduct extensive user experiments to start to 'map out' the outcomes in this set-up. Such user trials are time-consuming – for this reason we aim to capture aspects of user behaviour in interacting with the system

There will need to be a cycle of design, implementation, experimentation, and simulation. To achieve the stated goal this cycle will almost certainly have to be iterated several times (although during an assessment at most one iteration of this cycle would be possible).

The Hypothesised Structure

The appropriate structure for supporting the self-organisation of collective knowledge is the major research question of this project. It is thus impossible to fully predict what this will be before the research. However, based on previous work on the separate components of that will be used in this project, we can guess. The following represents our first goal and starting point from which to obtain data from the first user trial which will the enable a more informed design.

- The knowledge base will be initialised by taking an archive of relevant newsgroups and discussion lists, filtering them and giving them a preliminary semantic structure using pseudo-semantic algorithms based on word occurrence statistics (in addition to the thread, subject, date and author structure already present). This structure will consist of relevant links between nodes where a *weight* associated with each link represents the strength of the association between the nodes. In this way the knowledge base has something to offer browsers in the way of an information source with already more structure than an archive of posts.
- The knowledge base will be extended via short contributions from users of a limited number of types. These types will initially be just *questions* and *answers* but the choice of types will be expanded later to include such as *elaborations* and *rebuttals*. As it is contributed it will be linked to those nodes it refers to and the author of the nodes linked to will be automatically informed via e-mail to stimulate further contributions.
- When users browse the archive data about their browsing patterns will be captured and used to adapt the weights of the associative hypertext links following Hebbian reinforcement rules. In this way links which are more relevant will tend to be reinforced.
- When users browse the site those other nodes that are most relevant (as established by the associative reinforcement) will be suggested to the user in addition to the normal links in a special context-sensitive menu which frames the node.

User trial

Once a design has been implemented and tested as working a user trial is undertaken. It is this stage that is the most important. For although one can guess at appropriate designs for such a site, it is not possible to fully predict the effect of social interactions in a novel medium. The trials will be conducted around subjects with practical interest for a dispersed community of people. During the user trial of a design the maximum possible data will be collected, including: a detailed log of the browsing patterns of users, including the time of requests; a log of the nodes that are added; questionnaires to the users about their experience in using the system; in-depth discussions with individual users; the final structure of the knowledge base at the end of a period of interaction.

Simulation

Based on the data and information collected a simulation of the user interaction on the knowledge base will be constructed. This will aim to capture some important properties of user interaction with the system in an abstract way. The aim of this is to produce a dynamic description of the user trial that can be used in the exploration of conceptual and formal models of the user trials by the revelation of possibilities beyond the happenstance of the trials. It does not aim to be a predictive model, but rather an aid to enhancing the analysis of the user trial.

This will use techniques drawn from the field of agent-based social simulation as follows: *firstly*, the decision mechanisms reported by the users in interviews are abstracted; *secondly*, a relevant class of learning/decision algorithms are chosen which include these mechanisms; *thirdly*, these are implemented in artificial agents along with a model of the active knowledge base and tested to see if they would give the same distributions of results as those collected from the user trial, if not the parameters of the user models are adjusted until they produce as close a fit as possible; *lastly*, this model is analysed with respect to what it reveals as well as its short comings against the views of participants as well as the researchers expert opinion, if necessary previous stages are iterated.

Design of Next Version of the System

Traditional analyses of the results of the user trial, along with the comments of the users and the intuitions of the researchers involved are used to suggest improvements to the design of the system. These changes are they implemented in the model of the system in the simulation and tried out in the simulation of users to indicate some possible outcomes of the changes. This scenario generation will

allows us to make maximum use of the limited number of user trials possible.

Once the design of the next version of the system has been decided upon, the whole developmental cycle begins again. If the user trial has gone exceptionally well, in that it indicates that a workable system and simulation has been produced that does allow the significant emergence of self-organised knowledge, then some of the extensions or options listed below may be tried.

Design Extensions and Options

There are many possible extensions and alternatives to the suggested structure described above. These would be chosen and applied in response to the results of the user trials. There are far too many of them to be described in any detail here, but they include: mechanisms such as ranking for the explicit assignment of credit to user contributors to the knowledge base; mechanism to introduce weak tests of coherency or consistancy into the structure of nodes (based on algorithms taken from defeasible inheritance networks); introducing the possibility of node death; adjusting the amount of contextual information available to the users at any node; integrating global search tools and assessing their effect on the self-organisational properties of the knowledge base; implementing some weak inference engines on the structure to add in extra “organisational” nodes; and trying various mechanisms for the automatic notification of changes to users.

Implementation Details

The implementation of the adaptive web-site will be as a series of scripts called by the public domain “Apache” server software on top on the Linux operating system. The machine will be a high-end PC with fast network and storage facilities dedicated to the collective knowledge base. The agent-based social simulation will be implemented in one of the high-level agent-based or social simulation languages, for example: DESIRE, SDML or SWARM. These require fast PCs with a lot of memory.

Brief Survey of Existing Research

Learning and adaptive algorithms

…

Conferencing systems

…

Browsing tools

…

Component Technologies of SOCK

Evolutionary Algorithm

The dynamic as sketched above is basically evolutionary, stimulating both variation (creation of new answers) and selection or "survival of the fittest" (selective presentation of the most popular or authoritative answers). However, the interconnection of the ideas is still left to the individual authors, who have only a limited understanding of the context in which their contribution would fit. With an ever growing collection of nodes, that are constantly improved by their authors, it becomes more and more difficult for individuals to decide about the most relevant links from their pages. A number of algorithms have been developed by us that can tackle this problem, by letting the most appropriate linking patterns emerge from the collective choices made by all users of the network.

The basic principle is an extension of the rule of Hebb for learning in neural networks: concepts that are used together or in quick succession become more strongly linked. We have demonstrated the usefulness of such an approach for reorganising a website through a few, small scale experiments. The experiments used three learning rules: frequency, transitivity and symmetry. If a web user would move from page A to B and then to C, the frequency algorithm would increase the strength of the connections A -> B, and B -> C, the transitivity rule would increase A -> C, and symmetry would increase B -> A, and C-> B. We have shown that with these simple rules, a network of 150 nodes will self-organise quickly and efficiently from a random connection pattern to an associative network, where all related concepts are strongly linked to each other.

Recently, through theoretical reasoning, this approach has been generalised to a more sophisticated algorithm that in principle should reorganise a website more quickly and efficiently. The idea is that the time that a user spends reading a page (within certain limits, so as to avoid situations where a user opens a page and then forgets about it) is a good measure of that user's interest for the page. Let us

assume that a user spends quite some time reading page A, then browses quickly through pages B and C, then again reads D more attentively, and finally jumps quickly through E and F. In that case, we can assume that for someone who is interested in A, D is also quite relevant, but B, C, E, and F less so. In that case the algorithm will propose a strong connection from A to D, a weaker one from A to B and C, and an even weaker one from A to E and F.

The longer the delay between reading subsequent pages, the less we assume the pages to be mutually relevant, and therefore the weaker the strength their connection gets (decaying exponentially with the duration of the interval). However, the fact that two pages were read by the same user still implies that they have at least something in common and therefore the connection never becomes completely zero. This is an extension of the transitivity rule, which now is no longer limited to pages that are at most two steps away, but applies with decreasing force to pages that are an unlimited number of steps away from the starting point. The equivalent of the symmetry rule is simply to add a constant fraction of the connections strengths (e.g. A -> D) generated by this rule to the inverse connection D -> A.

The result of applying this algorithm to the web will be the generation of a square matrix whose elements are the strengths of the connections between any two nodes (corresponding to the rows and columns of the matrix) in the network. Every additional visit to the network will slightly change this matrix, so that it constantly adapts to the changing use patterns. The applications of this are manifold:

1 from the matrix a list of recommended links can be derived for any given page; these are the links with the highest strength, in the order of their strength.

2 by considering the mutual connection strength of links between nodes as a measure of similarity (or inversely, of distance), nodes can be clustered into groups of related topics. This can produce an automatic classification of subjects in the web, and generate a set of indexes for particular subdomains. Thus, new overall subjects or domains will spontaneously emerge from associations between different contributions.

3 by representing the interest profile of a particular user as a vector of activations (the activation of a node is proportional to the degree of interest by that user), individual recommendations can be generated for every user through "spreading activation": the activation vector is multiplied repeatedly with the connection matrix until the resulting product vector has stabilised.

Pseudo-semantic analysis using word occurrence matrices

The archive of posts on a subject is given a preliminary structure using the technique of Latent Semantic Indexing (LSI). In this technique each post is given a position in a high-dimensional space. The “axes” of this space are the relevant terms that occur in all the posts. The “co-ordinates” of each post is a function of the frequency of occurrences of these words in the posts. Then statistical techniques can discover those terms that tend to occur together (using the Eigenvectors of the Singular Value Decomposition). The initial structure of the set of posts can then be set in two ways: the decomposition can suggest additional “subject header nodes” (which are otherwise empty) to link to members of the clusters; and the strength of initial associative links can be determined inversely to a distance measure on the position of posts in the space.

Topologically Based Importance Statistics

Explicit or implicit evaluations have the drawback that not every visitor is similarly competent to judge about the quality of a page. The problem is that there is no objective way to determine who is most expert or authoritative on a particular subject. People become authorities or experts because they are considered as such by their peers. This implies a circular or bootstrapping mechanism: authorities get their status by being recognised by other authorities. Although this may seem paradoxical, recently a number of algorithms (PageRank and HITS) have been developed that solve exactly this kind of problem in the web: a webpage's authority is calculated recursively on the basis of the number and authority of other webpages that refer to it. Until now these methods have only been applied to the web at large, where the linking pattern is sparse and discrete, but mathematically they should work as well in a smaller website where the linking pattern is more fine-grained

Agent-Based Social Simulation

Recently the techniques of Distributed Artificial Intelligence and Multi-Agent Systems have been applied to the task of modelling social groups of humans in particular contexts. Such models are composed of a number of interacting agents, each of which is endowed with relevant cognitive abilities (e.g. planning, learning, or reasoning). These agents then “think” and interact in a way that directly analogous to the groups being modelled allowing the emergence outcomes to be explored. The cognitive abilities of the agents in the model are designed using a combination of theories of cognition taken from cognitive science and the decision making processes reported by the real actors. The results

of the interactions in the simulations are then compared to actual (or idealised) data and facts taken from the groups being studied. These models do not pretend to predict the actions of the groups being studied but rather are used to inform the researchers as to some of the possible outcomes and processes that can emerge, and thus inform their thinking on the subject. To support this activity various software tools and methodologies have started to emerge.

Deliverables

For *each* iteration of the developmental cycle there would be the following deliverables, made available on a publicly available web-site and through academic publication as appropriate:

1 A specification of the system.

2 The implemented software after it is debugged.

3 The data logs of the user trials.

4 The final state of the knowledge base.

5 The software code for the simulation of the user trial.

6 Design guidelines and directions gained from the experience and analyses of the trials.

References

...