Also in Proceedings of *Second International Workshop on Cooperative Information Agents,* eds. M. Klusch, G. Weiß, Lecture Notes in Artificial Intelligence 1435, Springer Verlag.

Agent Mediated Collaborative Web Page Filtering*

Shaw Green[1], Pádraig Cunningham[1], Fergal Somers[2]

[1]Department of Computer Science, Trinity College Dublin, Ireland.
[2]Broadcom Éireann Research Ltd., Dublin, Ireland
Shaw.Green@cs.tcd.ie, Padraig.Cunningham@tcd.ie, fs@broadcom.ie

Abstract. Intelligent filtering of multimedia documents such as World Wide Web (WWW) pages is an extremely difficult task to automate. However, the determination of a page's relevance to one's interests is a skill that comes with ease to most humans. This paper outlines a system architecture, developed using Agent Oriented Design (AOD), aimed at providing page filtering within a limited domain. This domain is specified via an explicit taxonomy. A prototype system was built using this architecture, which draws upon a user's innate ability to determine the relevance of web pages to their own information needs. It is argued that the resulting system incorporates the best aspects of existing Automated Collaborative Filtering (ACF) systems, whilst still retaining the benefits of the more traditional, feature based approach. In support of this claim and the general success of these systems, initial evaluation of the system is also reported.

1. Introduction

Many systems in recent times have attempted to tackle the difficult problem of selecting pages for presentation from the diverse and ever changing pool of material published on the World Wide Web (WWW). An important requirement of which is to only retrieve material that is of direct relevance to a user's interests and current information requirements.

It can be argued that these systems can be grouped into two major classes. First, there is the traditional feature based approach where the individual web pages are represented in some fashion by appropriate semantic structures or features. These features are then matched against other features that represent the user's interests and current goals. Examples of this type of system include the major search engines [URL1] which represent each of their indexed pages using a number of different feature based representations. These are then matched against the keywords that the user enters. These keywords assume a dual role of representing both the user's interests and their current goals. Another example of this class of system would be the sub-symbolic neural network news-group and web page filtering systems.

This approach contrasts with the featureless approach to the problem. Implementing a family of statistical clustering algorithms, these systems are often termed Automated Collaborative Filtering (ACF) [URL2] systems. Such systems are

* This research was sponsored by Broadcom Éireann Research Ltd.

essentially devoid of any form of "feature". Instead users of such systems are required to classify presented pages according to some personal measure of a pages "worth". This information is used to construct clusters of users with similar interests.

Both these approaches have their advantages and disadvantages. Feature based, particularly keyword based systems allow their users to focus the system not just on their interests but also on their current information needs. Any information outside this current subset of their interests will probably not be welcome, as it would be considered interesting but not relevant. ACF systems do not deal in short term requirements at all. Instead they focus on finding "Like Minded Individuals" and thus concentrate solely on interests. On the other hand ACF systems neatly side-step one of the major flaws of feature based systems, namely the poor representational capacity of features. Although keywords and other features are undoubtedly useful, they often fail to adequately represent the true meaning of the underlying text. This is because they lack the context that the full natural language text provides and are therefore often ambiguous.

This paper describes a system architecture designed to provide document filtering based upon a user profile within a specified domain. Details are presented of an implementation of our abstract architecture that allows web page filtering within an Irish context. This filtering system relies heavily on inter user collaboration. We claim that this system successfully combines the important characteristics of both traditional and Automated Collaborative Filtering(ACF) techniques

Section 2 of the paper describes other work relating to this topic. *Section 3* describes the architecture in some detail, whilst *Section 4* gives details of the example implementation of this architecture. *Section 5* gives details of some initial experiments conducted to assess the effectiveness of this system. The paper concludes with *Section 6* that presents the conclusions drawn from this work.

2. Related Work

When considering the existing work within this field, it is worthwhile to first consider the features common to all such systems. We identify an architecture consisting of three components arranged in a hierarchy. The topmost level of this hierarchy is concerned with the presentation of information to the user. There will therefore be information passed upwards from lower levels of the hierarchy as well as feedback from the user to pass to lower levels. The middle component in the architecture is concerned with selecting which material, from that available, to present to an individual user. This contrasts with the bottom-most level that represents those system elements that actually retrieve information from external sources and make it available for retrieval within the system.

The FAB system [Balabanovic 1997] consists of two major types of agent. Firstly, it has the concept of an agent for collecting material to provide to its users. These "Collection" agents obviously correspond to the information collection component of the architecture above. They retrieve material based on an agent profile that is then placed in a central information repository. The other major type of agent is the "selection agent" which corresponds to the information subscription layer of the reference architecture. These agents are responsible for drawing material from the central information repository for presentation to the user. Upon presentation of this information, the user ranks the material. This feedback is then used as the basis for modification of both the user's own personal profile and the profiles maintained by the collection agents.

Another architecture in the same domain is that proposed by Davies [Davies et al. 1996] as instantiated by their Jasper system. Each user in the Jasper Architecture is represented by an agent. When the user identifies a page as being of interest to them this agent is responsible for adding details of this page to that individual's store of interesting pages. These details include the URL of the page, any user annotations and a summary of the page's content, produced using a proprietary text summarisation package. Each agent also maintains an explicit profile entered by the user, which is supposed to adequately capture the user's information needs. The Jasper system as well as allowing a number of different querying styles, more interestingly also allows the communication of pages to other agents with similar interest profiles Jasper is an interesting system as it adopts the use of groups of users in an implicit rather than an explicit fashion. A possible weakness of this system however is its reliance solely on an explicitly entered user profile particularly as this profile again seems to be based on simple keywords.

Amalthaea [Moukas 1997] is another multi-agent system aimed at identifying pages of possible interest to a user based upon a profile of their interests. Again two major classes of agents are used, filtering agents and discovery agents. These agents are evolutionary in nature and are organised into a marketplace. That is the overall behaviour of the system emerges from individual agents in competition with one another based upon locally available information. In the case of Amalthaea the information available is in the form of user feedback on presented links. This information is "credited" between the information filtering agents that presented the link and discovery agent that retrieved it. Positive credit is assigned for pages the user rated highly. Correspondingly, negative credit is assigned for information presented that the user rated poorly. The information filtering agent, acts as a mask on the discovery agent's output, filtering documents based upon weighted keyword vectors.

SAIRE [Odubiyi 1997] is another multi-agent information retrieval engine. SAIRE operates in the space science domain, and adopts a somewhat different approach than those systems previously mentioned. SAIRE utilises legacy information retrieval systems and therefore concentrates on providing a scaleable architecture that is easy to use. SAIRE is organised into three levels. The topmost level contains those agents responsible for accepting input from the user. The system accepts input in a number of different modalities including written and spoken natural language. The middle layer of the architecture acts as a co-ordinator with the information retrieval engines at the bottom most level. One interesting aspect of the SAIRE project is the use of user stereotypes. Based upon previous work from the User Modelling community (e.g. ARCHON [Wittig 1992] and PROTUM[Vergara 1994]). SAIRE users assign themselves to one of a number of stereotypical user groups that are then specialised to fit the individual user. The SAIRE system is of particular interest to us as it makes use of stereotypical user profiles something we are interested in exploring.

To summarise therefore we have identified a three level architectural model within which a number of approaches are possible. We have described a number of recent systems that can be described in terms of this model and identified some of the benefits and weaknesses of these systems. In general, the more successful of these systems adopt some kind of market oriented approach. As with any marketplace it is populated by consumers and producers, the commodity in this case being information and the currency being user feedback. As always the success or failure of these systems lies in the details. In Figure 1 we identify a number of key characteristics that we use to classify and summarise the above systems. These characteristics will be returned to when we come to evaluate our own system. Some explanation of the "Implicitly" entries in the above table is required. Amalthaea does not explicitly maintain a user profile. In Amalthaea it is distributed amongst all the information filtering agents .If this representation is taken to be the user profile, then these are indeed adaptive. Again for Amalthaea the use of information filtering agents implies the existence of groups of users with similar interests otherwise an organisation with one filtering agent per user would make more sense.

Characteristic	**FAB**	**Jasper**	**Amalthaea**	**SAIRE**
Adaptive User Profile	Yes	No	Implicitly	Yes
Organised Credit Assignment	Yes	No	Yes	No
Support for User Groups	Yes	Partial	Implicitly	Yes
Support for Keyword Search	Yes	Yes	No	Yes
Support for Inter User Collaboration	No	Yes	No	No
Automatic Addition of New Information	Yes	No	Yes	No

Fig. 1. Summary of Key Features.

3. The ARC Architecture

The **A**utomated **R**ecommendation via **C**ollaboration (**ARC**) architecture is an attempt to provide a flexible platform upon which to build a variety of document retrieval solutions. The architecture is based upon an explicit user profile and a taxonomy to limit the scope of the system's domain of discourse.

As can be seen in Figure 2 below, any system based upon the ARC architecture will consist of a number of distinct components each responsible for different parts of the overall systems functionality.

Firstly an overall skeleton for the architecture is provided by a taxonomy agent. It maintains the categories under which documents are held and the inter-relationships between these categories. There is also a set of interest agents, one per category supplied in the taxonomy, which are responsible for managing all the content available within the system relating to the agents assigned category. Functions for this agent would including adding and removing items of content and of course supplying documents upon request matching the personal agent's requirements. Finally there will also be a set of personal agents, one per user. Each personal agent will be a fairly lightweight entity, responsible for allowing communication to, and receiving communication from relevant interest agents. Additionally each personal agent is responsible for maintaining some form of user profile detailing the user's interests.

As already noted above, this architecture describes a class of system that operates within a specified domain partitioned via an explicitly specified taxonomy. Although this is somewhat restrictive, it is not overwhelmingly so, as the categories can be made fairly broad, and the number of categories fairly large. This would result in a system that covers a broad range of possible content. There is also a presumption that some kind of profile will be available detailing the user's interests. Beyond these two presumptions the detail of any implementation of the architecture is undefined. This is a deliberate attempt to allow the widest range of possible systems to be based upon this architecture. Section 4, which follows, gives details of one such implementation of this architecture. The implementation provides a "What's Cool on the Net" service with an Irish flavour.

4. Implementation Details

This section details one such prototypical application. It tackles document filtration in the World Wide Web (WWW) domain. This domain, containing as it does, documents containing elements of many different media types, presents particular challenges when trying to determine overall document semantics. Also such a system has the advantage that it allows access to a large number of potential users via the Internet which would be impossible to obtain any other way. This is particularly important for systems, such as those developed based upon ARC, which depend heavily on inter-user collaboration.

4.1. Prototype Application

The Prototype application is designed to give two distinct views on the same set of information, namely links to web material with an Irish flavour. Which of the two views to be shown to an individual user at any given instance in time depends on the nature of their current goals. As was noted earlier, we distinguish between long term and short-term information needs. We support this distinction by allowing a keyword search style of interaction as well as a featureless view which functions as a kind of "What's cool in Ireland". Both these views share the same underlying data and are supported via the same set of agents. This system makes extensive use of recommendations from its user base rather than trying to use keyword based requests from existing search engines. The system therefore depends on a degree of altruism on the part of its users. As Webhound amongst other systems has demonstrated however, this is usually forthcoming providing the system proves genuinely useful to its user base from the outset. Thus a system such as this requires pump-priming with a number of web page links entered semi-manually by the system developer. The system was primed with around 2000 links spread across the 80 categories. The choice of these links is not of critical importance however, as the system acts to automatically filter out web links which are not found to be relevant. Further details of this implementation on an agent by agent basis are given below. The overall organisation of the implementation can be seen in Figure 2.

Taxonomy Agent

We initially envisaged a simple hierarchical taxonomy, this proved too simplistic however. This is due to a simple hierarchical structure not being rich enough to capture all the associations between interests that are naturally formed. The taxonomy provided in this implementation attempts to capture at least some of these relationships. The taxonomy provided allows the marking of pairs of interests as being related. Whilst maintaining the basic hierarchical structure, the taxonomy allows arbitrary relationships between pairs of interests to be constructed. These relationships between categories are based upon their mutual selection when the user first specifies their interests. The hypothesis in this case is that a user has an interest in both X and Y. If this is the case then topic X and Y may be of interest to others, who also specify either X or Y as being of interest. Of course one such user is not sufficient to establish this type of linkage and the agent allows for this by using a mechanism similar to that described for the weighting of web pages (See Below).

These linkages are utilised within the system to provide initial users with a stereotypical initial profile. This profile is an expansion of their explicitly stated interests. This profile can subsequently be pruned back automatically based on user feedback on presented pages.

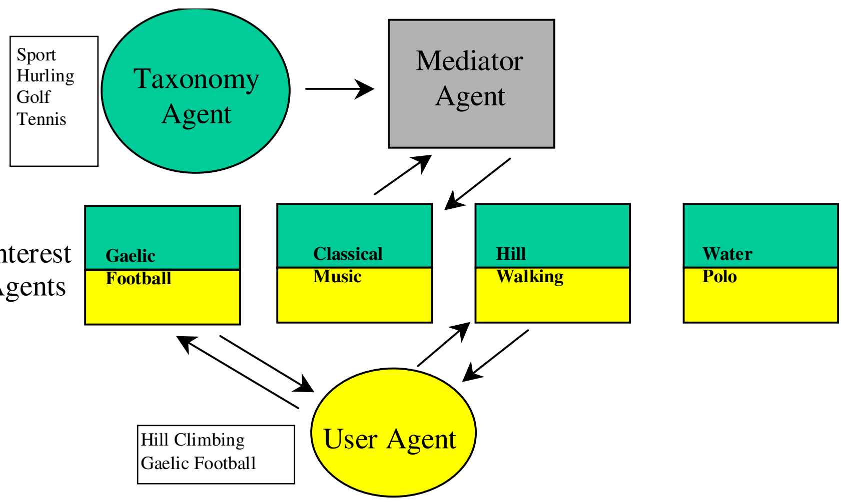

Fig. 2. Instantiation of the ARC Architecture

Interest Agents

These agents, 80 in this case, provide the major functionality of the system. As previously noted the user's interests can be split into long and short-term requirements. Clearly the long term requirements can be given adequate expression via the user profile maintained by the user’s personal agent (see below). What about the short term goals however? This system gives the user an ability to communicate these by allowing them to enter a set of keywords, which characterise the web pages they are currently seeking. Note that these keywords are utilised within the context of the user profile, which it is thought should considerably reduce the problems associated with ambiguity of keywords. Consider for example a word with multiple meanings such as “club” which for example could stand for such things as football clubs, dance clubs & golf clubs depending on the context. If a user of our system has flagged an interest only in football and not in golf or dancing then this keyword is no longer ambiguous. This is, it is felt, one of the major advantages of the use of a user profile in this type of system. In order to support this use of keywords, the interest agents are responsible for creating a keyword vector representation of each page entered into the system. When the user presents their list of keywords to the system these are simply matched against the keywords for each page.

As previously mentioned the interest agents also have a mechanism for removing irrelevant or poor quality material, each time a page is presented to a user, they are asked to rate its relevance to them. This information is used to update the page's weight, decreasing it if the page is rated poorly, increasing it for a good rating. A page's weight also decays naturally if it is not presented at all. Any page's weight that falls below a pre-defined threshold is removed from the system. New documents are introduced into the system by means of user recommendations. In this case the user is asked to submit the URL along with a title for the page and a measure of its usefulness, this measure is used to set the documents initial weight.

The featureless style of interaction presents different problems. We make use of this mode of interaction to allow for the adaptation of the user profile as the focused keyword style of interaction is simply too focused for this type of adaptation to take place. When the user interacts with the system in this mode, they are presented with a set of links based loosely on their user profile. In fact the group from which each link is drawn is decided using a Genetic Algorithm(GA) style roulette wheel. Eighty percent of the wheel is allocated to the groups within the user profile according to their allotted strengths. The other twenty percent is split equally between the remaining groups in the taxonomy. Furthermore in order to decide which link from within each interest group to select a further roulette wheel is used. Each link within the group is allocated a slot the size of which is determined by the strength associated with the link. These strengths are updated as indicated in our discussion of the keyword-based view. The total number of links presented is dependent on the user’s preference settings.

Personal Agents

The personal agents are fairly lightweight being concerned chiefly with facilitation of communication between each individual user and the rest of the system.

The main function of this agent is to maintain the user profile. This requires the agent to update the strength of each element within the users profile based upon feedback given on items presented drawn from that category. Items presented from outside the profile and ranked highly cause the corresponding category to be added to the profile at an initially low strength.

The results of an initial evaluation of this prototype can be found in Section 5 of this paper.

5. Experimental Evaluation

This section gives details of two experiments conducted in order to empirically demonstrate the effectiveness of the various learning mechanisms used within the prototype system. The first experiment was very tightly controlled and was constructed in order to demonstrate the effectiveness of the filtering mechanism within a single interest category. The second experiment allowed the user to use the system in a less constrained manner and was intended to demonstrate the systems ability to refine a user profile.

5.1. Interest Group Filtering (Experiment 1)

Experimental Procedure

As was previously mentioned this experiment was very strictly controlled and was designed to show the effectiveness of the interest group based filtering of web pages. The subjects, eight in total were drawn from the Computer Science department here in Trinity College Dublin. Each subject was asked to imagine they were planning a short break into a particular area of Ireland (Wicklow) and were planning to engage in a number of outdoor activities in this area. It should be noted that all subjects were native to Dublin and thus had at least some familiarity with the area concerned. Each experimental subject was assigned the same single interest profile for the purposes of the experiment. The experiment itself consisted of each subject carrying out the following three simple tasks :-

- First, each subject was requested to enter a three keyword query which was given to them as part of the experiment. This three keyword query was the same for all subjects. Entering this query resulted in a set of documents being returned which they were requested to rate with respect to relevance.
- Second, each subject was requested to provide to the system two pages which they had not been provided with, but which they felt would be relevant if they were planning such a holiday. Obviously, some subjects would have this material to hand whilst for others this would require some searching.
- Finally, after everyone had completed stage 2 everyone was requested to redo the first stage. This resulted in a new list of material which they were again requested to rate.

Results

The results of this experiment are presented in graphical form as Figure 3. They clearly show a marked and significant difference between the acceptability of the initially presented set of data and that presented after the filtering process has taken place. This is clear, even though the system has only been in operation for one iteration. If we average the results across users and take the mean value and associated standard deviation this result is equally clear. We get a mean of 1.4 with a standard deviation of 4.5 with the initially presented data This indicates a low acceptance rate for the material although the level of acceptance varied markedly as indicated by the standard deviation. After the filtering process has taken place the acceptance value rises to 3.1 and the standard deviation drops to around three. Both these figures are pleasing as they indicate both a rise on average in acceptability along with a convergence across the user group.

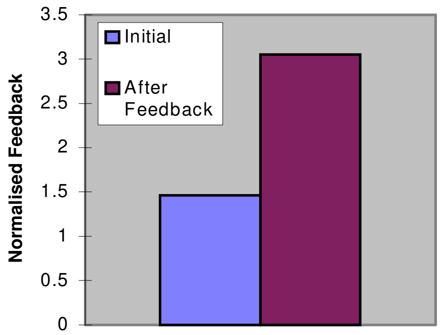

Fig. 3. User feedback before and after learning.

5.2. Adapting to fit an individual user (Experiment 2)

As well as demonstrating the effectiveness of the page filtering, we also hoped to demonstrate the effectiveness of two aspects of the system which concern themselves with user adaptation.

First, one would like to show that the process of modifying the initially entered user profile is useful. That is, that the initially entered user profile isn't the optimal profile for the user concerned. Second, one would also like to show that the system view of the usefulness of a particular document in a given context mirrors that of its user base. It is with these two points in mind that the following experiment was devised.

Experimental Procedure

Twelve subjects were requested to create themselves an initial interest profile from the categories of interest available within the system. They were then requested to download and subsequently rank a set of pages. They were then requested to download and rank a second set of links which would be based upon their updated profile.

Results

Figures 4 and 5present in graphical form the results obtained from the experiment described above. As can be seen from Figure 4 the second ranking of pages was consistently better than the first indicating that the system does indeed learn to better estimate the interests of the user. This effect was however somewhat smaller than we had hoped - a point which is discussed in our conclusions (Section 6).

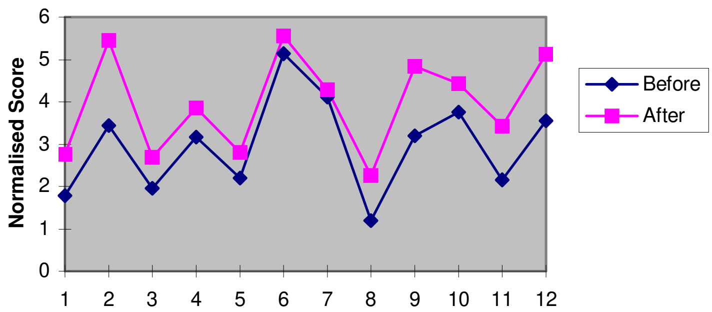

Fig. 4. Analysis of User feedback.

Figure 5 is intended to illustrate the narrowing of the gap between the systems expectations of a users perception of a pages quality, and the reality. Unfortunately the picture is somewhat clouded by a separate process illustrated in Figure 4 i.e. an overall increase in the quality of the pages presented to the user. Even taking this factor into consideration it would seem fair to state that some convergence between system and user is occurring.

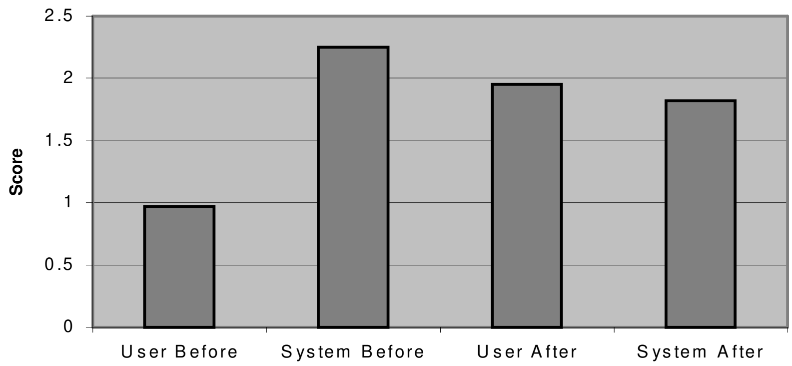

Fig. 5. Comparison of System and User Page Ratings.

6. Discussion of Results and Conclusion.

As was mentioned in Section 2 we have identified a number of characteristics, which can be used to distinguish between different agent based information retrieval solutions.

The results presented above provide guidelines as to the direction to take on some of these issues. For instance in a domain such as ours where the topics of interest are simple and clearly defined it could be argued that the effort required in building a system which constructs a dynamic profile is unnecessary. The number of new interests added to a user profile after the user initially specifies their profile is typically very low indicating that the initially selected interests are satisfactory. The explicit nature of our profiles is potentially useful however as the user could be shown their profile at any time and make modifications to it if they are unhappy.

Similarly the facility for users to recommend material to other like minded users is clearly useful. However in the absence of a large user base (as was the case for the system described here) this needs to be augmented with automatic page retrieval facilities.

The use of the user profile and associated interest agents to provide context for a keyword based search is also of interest and warrants further investigation.

In conclusion, it is felt that the prototype system was successful in its goals of bringing together feature and featureless modes of interaction within the one system. Such a system provides an interesting platform for the investigation of issues relating to information retrieval and user modelling. The results of our initial evaluation show that the system was effective in its provision of pages. It would however be interesting to apply the same basic architecture to the problem of sensibly expanding on a keyword search thereby leveraging the power of the pre-existing search engine technology.

7. References

URL 1, Altavista Search Engine Web Page http://www.altavista.digital.com

URL 2, Collaborative Filtering Page http://www.sims.berkeley.edu/resources/collab/

Balabanovic M. An Adaptive Web Page Recommendation Service. In Proceedings of Autonomous Agents, Marina del Rey CA, USA 1997

Davies N.J., Weeks R. & Revett M.C. Information Agents for the World Wide Web in the BT Technical Journal 14:4 pg 105-123. 1996

Moukas A. and G. Zacharia. Evolving a Multi-Agent Information Filtering Solution in Amalthaea. In Proceedings of Autonomous Agents, Marina del Rey CA,USA 1997

Odubiyi J. et al. SAIRE - A Scalable Agent-based Information Retrieval Engine. In Proceedings of Autonomous Agents, Marina del Rey CA, USA. 1997.

Wittig T. (editor) ARCHON: An Architecture for Multi-agent Systems. Ellis Horwood, 1992

Vergara H. PROTUM: A Prolog-based tool for User Modelling . University of Konstanz, D-78434, Konstanz, Germany 1994.