# INTELLIGENT AGENTS GENERATING PERSONAL NEWSPAPERS

D. CORDERO, P. ROLDAN, S. SCHIAFINO, A. AMANDI

*ISISTAN Research Institute, Facultad de Ciencias Exactas,*
*Universidad Nacional del Centro de la Pcia. de Buenos Aires*
*Campus Paraje Arroyo Seco - (7000) Tandil - Bs. As., Argentina*
*email: amandi@exa.unicen.edu.ar*

Keywords: Internet Agents, Intelligent Agents.

Abstract: NewsAgent is an intelligent type of agent that has the capability of generating personal newspapers from particular user preferences extracted by observation and feedback. This agent generates personal newspapers using static word analysis for extracting a global classification and case-based reasoning for dynamic sub-classification. The agent observes users by an applet with capabilities of detecting changes of pages. It also records the routine of reading newspapers of each user for analyzing readings in terms of their routines. The contributions of this work are both a software architecture for interface agents moving on web pages and the classification of specific themes using case-based reasoning.

## 1. INTRODUCTION

Big amount of data is today available in the world from web pages. Sometimes we spend our time looking for relevant data for us. In this activity we use tools that search pages from a sequence of words. Generally we find a lot of irrelevant pages among interesting pages as result. From this common situation, is is observable one important problem: the spending of our time in such searching.

For avoiding such losing of time we can built an intelligent agent that execute o that activity per us. Such agent could search relevant data for us and then suggest a small set of interesting pages for reading.

Aiming such agent, several problems must be resolved. How to collect the user interests, how to detect alterations in such user interests, how to match data for achieving relevant pages and how to assist users without become heavy with too help are some of those problems.

All this being said, it is worthwhile to point out the building of intelligent assistants. Until now, several intelligent agents have been developed (i.e. (Lieberman, 1995; Jaczynski, 1997; Joachims, 1997; Moukas, 1998)). These experiences show the necessity of that type of agent and that we have a lot of work for solving the limitation of the actual agents.

In this paper we present an intelligent agent that assist users in the navigation of news framed in newspapers, generating a personal newspaper. We take the themes involved in the newspaper world as study domain because those use changing information on specific news sites. Our agent works on a conventional

Web browser such as Netscape, which is usually used for reading electronic newspapers.

For clarity of presentation, the introduction of our agent named NewsAgent is made from two perspectives. The first treats on capabilities of this type of agent for assisting users. The second exposes how the agent goes knowing about user preferences and how detects pages related to such preferences. In resume, the structure of the paper is the following: Section 2 the agent functionality is introduced. In Section 3 the agent architecture is presented. Then, related works and conclusions are presented.

## 2. THE AGENT NEWSAGENT

We usually spend part of our time reading news from several newspapers. But generally our attention is concentrate on particular interesting subjects. If we could find a newspaper with only date related to our interesting subjects we could reduce our time of analyzing a lot of information.

NewsAgent is a type of agent that is able for observing users when they are reading electronic newspapers and it has the capability of deducing the interesting subjects of particular users. Moreover, this agent enables users explaining their interests, but it is an optional way for informing only clear interests to personal agents.

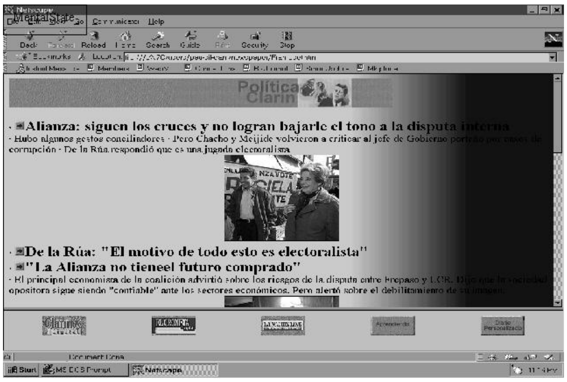
For each user, a profile is built analyzing their read pages. The subjects involved in the pages are extracted analyzing the address, the HTML code and the text of the body of each one of them. In the text analysis, case based reasoning is used, where a reading represents a case.

The interesting subjects are internally sorted considering the degree of user interest of each one of them. This interest degree is mainly obtained by time control. Users reading a page during a given range of time show their degree of interest in the subjects inside that page.

Basing basically on those dates, NewsAgent builds a personal newspaper for one particular user. Notes of several newspapers compose this personal newspaper and it is presented in such way that the more interesting subjects are located in preferential places on the main page. In such presentation, the agent uses text and images from original newspapers. Figure 1 shows a personal newspaper, which is composed by notes of newspapers written in Spanish. Each user defines both the idioms and the newspapers that the agent will use for composing its personal newspaper.

After that, NewsAgent tests the resulting personal newspapers. Users are observed when they are reading their personal newspapers. The time consumed in each note is analyzed in terms of their routine for reading newspapers. For example, if a user usually spends one hour reading news from newspapers and during a range of one hour a supposed interesting note presented in its personal newspaper is not read, the agent reduces the certainty degree of its supposed affirmation. Additionally, the user can indicate the certainty about the suggestions of the agent. With all this information, the mental state of user is altered for the next newspaper generations.

Fig. 1. A view of a personal newspaper.



Alianza: siguen los cruces y no logran bajarle el tono a la disputa interna

· Hubo algunos gestos conciliadores · Pero Chacho y Meijide volvieron a criticar al jefe de Gobierno porteño por casos de corrupción · De la Rúa respondió que es una jugada electoralista

De la Rúa: "El motivo de todo esto es electoralista"

"La Alianza no tieneel futuro comprado"

· El principal economista de la coalición advirtió sobre los riesgos de la disputa entre Frepaso y UCR. Dijo que la sociedad opositora sigue siendo "confiable" ante los sectores económicos. Pero alertó sobre el debilitamiento de su imagen.

## 3. THE ARCHITECTURE

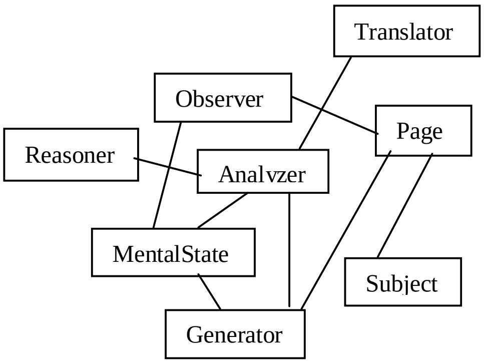
The architecture of NewsAgents is mainly composed by eight basic components: Page, Subject, Observer, Mental State, Analyzer, Translator, Reasoner and Generator. Figure 2 shows these components and the relationships among them.

The Page component represents particular Netscape pages. An instance records the structure of one page and the subjects that it contains. The Subject component records the characteristics of a particular subject.

The Observer component is an important part of the agent. This part has the ability of moving inside the network. In our implementation, applets take this role.

The Mental State component is responsible for recording interesting subjects, preferences, routines and particular experiences related to the navigation. This information is used by several related components as we can see in Figure 2.

Fig. 2. A static view of the architecture.



The Analyzer component collaborates in the building of the mental state of each user and also in the definition of the subjects of newspaper notes. This component uses other several components as the translator, the mental state and the reasoner.

The Translator component has the capability of translate words and sequence of words in several idioms. This component enables the generation of personal newspapers in several selected idioms.

The Reasoner component is responsible of managing the semantic of the notes. For doing so, it uses a case-based reasoner. Through this reasoner, punctual themes can be learned, which sometimes can be incorporated in the subject tree by the analyzer component.

The Generator component has the needed behavior for enabling the generation of a personal newspaper. The analyzer and the mental state are the components that do part of this work.

The next sections expose details of the more relevant components, the observer, the analyzer and the reasoner.

## 3.1. The Observer and The Analyzer

For building a personal newspaper for a given user it is necessary to collect information about its interests, and thus to built its mental state. For collecting interests, our agent observes this particular user reading electronic newspapers.

Recording certain characteristics of those pages it is possible to deduce the themes preferred by the user. Among the characteristics that we want to record from a given page we can mention the origin newspaper, the section, the specific theme of note and time used in its reading.

For analyzing the collected information, we use applets Java with appropriate methods for achieving this goal. Those applets, as part of a HTML page, are combined with functions defined in JavaScript which enable the information access from HTML pages.

The definition most popular of applet says that an applet is a small (let) application (app) Java accessible from an Internet server, transportable in a network, automatically installable, and executable as part of a Web page. An applet is a short application because it has to move inside the network. An applet is based on a graphic format without any independent representation. In other works, it is a component to be embedded in other applications.

Applets are generally executed inside Web pages through a browser like Netscape, HotJava or Microsoft Explorer. It is not necessary to take into account the existence of a main method, an applet assumes that the code is being executed inside a browser.

Opening a HTML page containing an applet, an instance of the applet is allocated, and then its execution starts. When the user leaves a page, following a link as example, the applet stops the execution, which is restarted when the user returns to it. Finally, when the user finishes the execution of the browser, the applet is stopped and the allocated resources are released before of leaving the browser.

For achieving data about user interests, our agent observes the way that the user followed on the net composed by pages of electronic newspapers. The analysis process of those collected data is made in the same moment in that the user is reading the news pages. This goal is achieved by the interception of actions taken by users, without any interference of their reading.

The agent NewsAgent observes to the user reading given newspapers in HTML pages. The data analysis is made by Java applets, which are inside this page. Users select newspapers and their agents observe and analyze their reading. A function JavaScript is responsible for collecting the URLs visited, including the time used in each one of them.

For detecting when a user changes of page, it is permanently controlled any alteration in the current URL. This process depends on the electronic newspaper currently being read, in term of each newspaper uses a different frame structure.

The observation of changes in the URL implies to take the execution control of the browser, avoiding thus that the user follows any link. Therefore, the agents take the current URL each five seconds, using a JavaScript function. This function called setInterval uses as parameters a function name and a given time in milliseconds. It invokes the function specified in the first parameter each certain interval of time specified in the second parameter.

The function setInterval() invokes a JavaScript function named mainF() each five seconds. This time was considered adequate for detecting whether the user changes of page enabling also a normal reading of the newspaper.

By the mainF function, the agent obtains the URL of the current page from the location attribute of the current frame. It is compared with the URL recorded in the previous invocation of the function. If they are different, the agent concludes that the user is reading a new page, which has to be analyzed.

For each page visited for the user, the agent records the time in which it is detected the change of URL and the URL of the current page. Then, the methods related to the observer applet are invoked. The more relevant methods are *beginVisit*, which treats the start of the visit to a page and *endVisit*, which it is invoked when the visit finishes.

The method beginVisit() loads a copy of the page being read by the user on the local disk. Then, the user analysis is made from both the HTML code of the pages and the URL itself.

From the URL, the agent obtains the newspaper name and the name of the section that belong to. This is possible because these data includes relevant words or codes that enable their identification. For instance, from the URL http://www.*clarin*.com.ar/diario/98-10-20/*pol*_sum.htm, we can detect that the page treats about the politic of the Clarín newspaper.

From HTML code of the page, the agent extracts the page title, the text of the note, the hyperlinks to other pages and the images. For doing so, the agent analyzes the specific tags that describe each one de those elements. Also, the agent discovers the theme involved in the page from the text itself.

Meanwhile the agent analyzes the page being read, the method endVisit() establishes the time of ending of the reading of the current page. From here, a number representing the time of reading is calculated by the difference of the time of ending and of starting of the visit, considering specially a time for loading the page. With this information, the agent selects only those pages that he supposes of interests for the user.

In resuming, the observer and the analyzer of the agent built their mental state in which the model of the user is the main objective.

The next section exposes details of the analyzer related to the usage of experiences of reading of particular users.

## 3.2. The Reasoner

Case-based reasoning (Kolodner, 1993) is a technique that have been used in several areas, including information retrieval (Lenz, 1998).

Experiences of a user reading notes of newspapers can be used for detecting interesting themes for him. Each reading is recorded as a case. The case base is used for detecting too specific interests from similarity and frequency analysis of given types of readings.

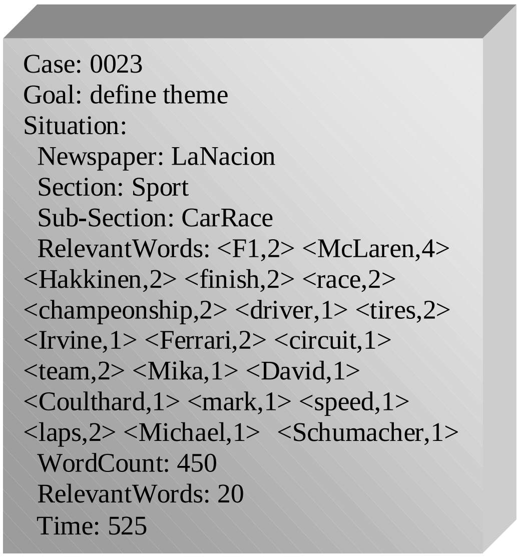
For example, the fact that a user is reading a note of international politic involving Spain and Italy say to the agent that that user has a possible interest in general international politic and/or politic facts involving Spain and/or politic facts involving Italy. This isolated fact gives us few data about the interests of the user, but if we analyze several readings we can detect the real objective of the user.

The analyzer component of the agent uses a case-based reasoner for discovering user interests from specific readings. Each note that the user has read it is recorded together with the time used for its reading. For instance, consider that the user has read a note about the big differences among the McLaren team and another teams. Mika Hakkinen and David Coulthard are highlighting their leader roles in the championship from that team. This note was read in the LaNacion newspaper site, specifically in the sport section. Part of the resulting case is showed in Figure 3.

The situation of these cases is defined from words found in the text of the note. The solution of that case is the definition of the theme.

Each situation of a case is composed by relevant words. Taking both cap words and nouns, the agent collects this set of words. The cap words are considered relevant elements for identifying a note because they represent people names, counties, cities, companies, and so on. The nouns are considered relevant words because they define the context in which the cap words are used.

Fig. 3 - A part of a case.



Case: 0023
Goal: define theme
Situation:
Newspaper: LaNacion
Section: Sport
Sub-Section: CarRace
RelevantWords: <F1,2> <McLaren,4> <Hakkinen,2> <finish,2> <race,2> <champeonship,2> <driver,1> <tires,2> <Irvine,1> <Ferrari,2> <circuit,1> <team,2> <Mika,1> <David,1> <Coulthard,1> <mark,1> <speed,1> <laps,2> <Michael,1> <Schumacher,1>
WordCount: 450
RelevantWords: 20
Time: 525

For comparing cases, a number is associated with each word. This number indicates the weight of the word in the definition of the topic of the note. Then, the words and the associated weights specify the definition of the topic of that note.

The section of a note defines a general topic of the note. For detecting specific topics, the agent uses case-base reasoning. For doing so, the reasoner plays with the weights.

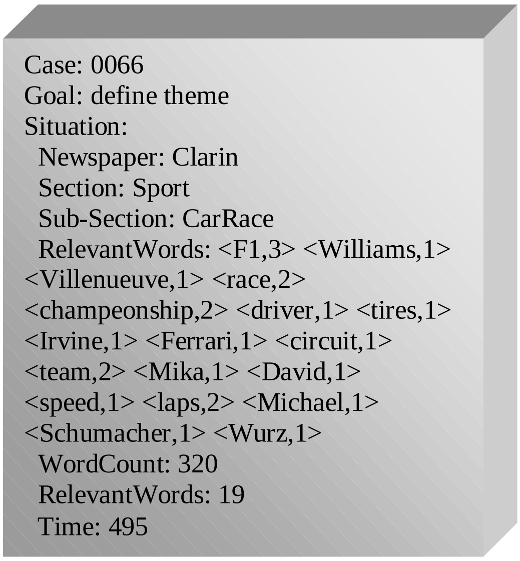
The weights of each word can change with the next readings. The changes reflect the interests of the user reading the news. Thus, for example, a more approximate notion of the topic of interesting of the mentioned note about international politic is achieved. This approximation is made when the user reads, for example, new notes about international politic involving Italy and different countries of Spain. Thus, the weight of Italy grows and the weight of Spain downs.

Thus, when the weights become stable numbers the new specific interesting topics are included as sub-topics of the old section topics. In the example, the new topic will be a subtopic of existing international politic in the topic tree.

Until here, we expose the use of case-based reasoning for defining specific themes. Now, we will present the application of case base in the detection of the topic of a note, in a general or specific way.

Therefore, cases are also used for identifying the theme of a note, which can be a general theme defined in the topic tree or a specific theme discovered by case-based reasoning technique. For example, Figure 4 presents a case resulting of a note of the Clarin newspaper, which treats a fact of sport, specifically F1. The agent needs to decide if this note is interesting for a user. Then, it is compared with previous reading cases of supposed interests of the user.

Fig. 4 - A new case.



```
Case: 0066
Goal: define theme
Situation:
  Newspaper: Clarin
  Section: Sport
  Sub-Section: CarRace
  RelevantWords: <F1,3> <Williams,1>
<Villenueuve,1> <race,2>
<champeonship,2> <driver,1> <tires,1>
<Irvine,1> <Ferrari,1> <circuit,1>
<team,2> <Mika,1> <David,1>
<speed,1> <laps,2> <Michael,1>
<Schumacher,1> <Wurz,1>
  WordCount: 320
  RelevantWords: 19
  Time: 495
```

The agent uses the case-based reasoner for discovering the more similar previous cases and thus to know the topic treated in the analyzed newspaper note. For doing so, a number is calculated indicating the degree of similarities between the new case and one specific case in the case base.

Calculating the degree of similarity between two cases, the following formulae is used:

$$
\begin{aligned}
DS(C^1,C^2) = \\
&(P^1 * Sim_n(N^1,N^2) \\
&+ P^2 * Sim_s(S^1,S^2) \\
&+ \textstyle\sum^n_{i=1} (P^3_i * Sim_w(W^1_i,W^2))) \\
&\qquad\qquad\qquad / (n+2)
\end{aligned}
$$

where DS is the similarity function, which uses two cases as arguments. The weight of the newspaper is $P^1$, being $N^1$ and $N^2$ the newspapers of the notes represented in the cases $C^1$ and $C^2$ respectively. The weight of the section is $P^2$, being $S^1$ and $S^2$ the newspaper sections of the notes represented in the cases $C^1$ and $C^2$ respectively. The symbol $P^3{}_i$, being $W^1{}_i$ and $W^2$ words belonging to the notes of the cases $C^1$ and $C^2$, indicates the weight of each word. Finally, *Sim* is the similarity function between two simple descriptors and n is the amount of words of the new case.

The similarity function $Sim_n(N^1,N^2)$ returns the degree of similarity of two newspapers. The similarity between two newspapers is made using a classification of the orientation of each one of them. This degree of similarity is qualified by a weight of the importance of this descriptor.

The function named $Sim_s(S^1,S^2)$ returns the degree of similarity of two sections of two newspapers. For doing so, the distance between these two section names is calculated using a tree of section subject.

The similarity function $Sim_w(W^1{}_i,W^2)$ calculates the degree of similarity among one word of the new case and some word or words of the old case. Calculating of the similarities of the texts, n invocations to this function are made. The number n is the quantity of words of the new case. Each work of the new case is compared with each of the old case, considering the quantity of occurrences of each one.

Coming back to the example about sport, the case numbered 0066 has the same theme of the case as 0023 as identification number. This is the result of the application of case-base reasoning on the reading cases.

That result is obtained applying the mentioned similarity formulae on several cases of the case base.

The newspapers of each note are different, but this is not a significant descriptor, which it is reflected with a lightweight. The section of each note is the same, then the match is complete.

The text is similar. The degree of similarity is calculated using the $\mathrm{Sim}_w$ function. Thus, each word of the case 0066 is compared with the words of the case s of the case base. When the new case is compared with the case 0023, full similarities such as one occurrence of the word Schumacher in both notes, and nearly quantity of occurrences of the word Ferrari. The rest of the words are compared using a similarity tree specific for sport. In this way, the word McLaren matches the word Williams. The matching is not full, but the match degree is high because both words refer to F1 teams.

One point that is still not clear is the order of matching of the similarity function among words. For instance, the word Williams of the new case can be matched with both words McLaren and Ferrari of the old case.

Matching words, the case-based reasoner takes the best match of each word. It also removes the words used in this process from the list of possible matching word. Thus, the rest of the application of the similarity function uses different words.

In the example, the best matching of Ferrari is the word Ferrari, being Ferrari disposed for using in the rest of the matching process. Thus, the word Williams matches with the word MacLaren, being it the only option for the similarity function.

In resume, this section has exposed the usage of case-based reasoning in the definition and detection of subject notes. This functionality is combined with the information obtained by the observer and treated by the analyzer. The case-based reasoner is indeed a part of the analyzer component, which is defined in a separate way for designing a flexible component. Thus, this component can be easily adapted or dynamically changed.

## 4. RELATED WORK

A lot of assistants of web browsers have been built in the last years (Lieberman, 1995; Etzioni, 1996; Ackerman, 1996; Jaczynski, 1997; Joachims, 1997; Moukas, 1998; Terveen, 1998) because the amount of information is constantly growing.

The existing assistants can be classified in terms of their objectives. To search of specific topics, to search general topics, to control changes in specific pages, to build personal pages are some of the most common objectives, being the last the description more approximate to our work.

Among the assistants that generates personal pages for given users, we can mention to Letizia (Lieberman, 1995). But using techniques such as case-based reasoning, we have to mention to Broadway (Jaczynski, 1997).

All those assistants have usefulness in different moments in terms of the goals of each moment, but all of them are used for each one of us. An integration of these tools will become a necessity in a short time. The work presented here is the first step towards such integration. The choice of a changing domain as the news world was a decision motivated by the objective of analyzing several problematic alterations. These studies enable us to detect changes in both existing pages and navigation routes and also to treat stable and no stable application domains.

## 5. CONCLUSIONS

In this paper we have been presented an intelligent agent -NewsAgent- assisting the generation of personal electronic newspapers from the interesting topics of each user. Our study domain is the actual world news framed in electronic newspapers because it treats specific themes on changing pages.

The current version of NewsAgent is implemented using basically Java and JavaScript for adding extra functionality to a web browser. Javalog (Amandi, 1998), an integration of Java and Prolog, is used for managing logic knowledge in a Java context.

The agent is being used for different users and the few tests made that the suggestions are coherent with the preference model of our users.

The contributions of our research are the architecture for interface agents working on web pages and the classification of themes using case-based reasoning.

This project is being currently continued without to put constraints in the information domain.

## REFERENCES

Amandi, A.; Iturregui, R. & Zunino, A., 1998, Object-agent oriented programming, in proceeding of the Argentine Symposium on Object Orientation, Buenos Aires.

Ackerman, M.; Billsus, D.; Gaffney, S; et. al. 1996, Learning Probabilistic User Profiles: Applications to Finding Interesting Web Sites, Notifying Users of Relevant Changes to Web Pages, and Locating Grant Opportunies. AI Magazine 18-2: 47-56.

Etzioni, O. 1996, Moving Up the Information Food Chain: Deploying Softbots on the World Wide Web, in proceeding of the AAAI96.

Jaczynski, M. & Trousse, B. 1997, Broadway: A World Wide Browsing Advisor Reusing Past Navigations from a Group of Users, in proceedings of the Third UK Case-Based Reasoning Workshop, Manchester.

Joachims, R.; Freitag, D. & Mitchell, T. 1997, WebWatcher: A Tour Guide for the World Wide Web, in proceedings of the Fifteenth International Joint Conference on Artificial Intelligence.

Kolodner, J. 1993, Case-Based Reasoning.

Lenz, M.; Hübner, A. & Kunze, M. 1998, Question Answering with Textual CBR, in proceeding of the International Conference on Flexible Query Answering Systems, Denmark.

Lieberman, H. 1995, Letizia: An agent that assists web browsing, in proceedings of the International Joint Conference on Artificial Intelligence, Montreal, August.

Moukas, A. & Maes, P. 1998, Amalthaea: An Evolving Multi-Agent Information Filtering and Discovery System for the WWW, Autonomous Agents and Multi-Agent Systems, v.1, n.1, 59-88.

Terveen, L.; Hill, W. & Amento, B. 1998, Collaborative Filtering to Locate, Comprehend, and Organize Collections of Web Sites, SigArt Bulletin, v.9, n.3&4, Winter, 10-17.