Evaluating Adaptive User Profiles for News Classification

Ricardo Carreira Jaime M. Crato
Instituto Superior Técnico
Av. Rovisco Pais, 1000 Lisboa, Portugal
{rjlnc,jmc}@mega.ist.utl.pt

Daniel Gonçalves Joaquim A Jorge
Computer Science Department, IST
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
djvg@gia.ist.utl.pt, jorgej@acm.org

ABSTRACT
Never before have so many information sources been available. Most are accessible on-line and some exist on the Internet alone. However, this large information quantity makes interesting articles hard to find. Modern Personal Digital Assistants (PDAs), mobile phones, and the advent of ubiquitous computing will further complicate matters. Away from the desktop, the time to select important articles might be even harder to find. Strategies to select relevant information are sorely needed.

One such strategy is content-based filtering, coupled with User Profiles. Our prototype uses a Bayesian classifier to select articles of interest to a specific user, according to his profile. The articles are extracted from web pages and displayed in a zoomable interface-based browser on a PDA. Interests may change over time, making it important to keep the profile up to date. The system monitors the users' reading behaviors, from which it infers their interest in particular articles and updates the profile accordingly. Results show that, from the start, most articles are correctly classified. An initial profile opposite to the user's actual interests can be reversed in less than ten days, showing the robustness of our approach. A user's interest in an article is inferred with a high degree of accuracy (over 90%).

Categories and Subject Descriptors
H.3.3 **[Information Storage and Retrieval]**: Information Search and Retrieval – *Information filtering*. H.5.2 **[Information Interfaces and Presentation]**: User Interfaces – *user-centered design*.

General Terms
Algorithms, Human Factors.

Keywords
User profiles, Content-Based filter, Bayesian classifier, mobile and ubiquitous computing.

1. INTRODUCTION
Many technological breakthroughs in recent years provide people with almost immediate access to information. Its sources on the Internet have grown both in number and diversity. Currently, mobile phones, some Personal Digital Assistants (PDAs), and other wireless devices allow information to be accessed almost anywhere, at any time. The advent of ubiquitous computing will make access to those sources even easier. All these information sources make it hard for users to choose the most adequate according to their interests. Finding useful information of personal interest has become difficult for a regular Internet user. Ideally, users should be able to take advantage of the wide range of available information while being able to find only that which is interesting to them, a small fraction of the total

Manually searching and analyzing the available news articles to select those considered interesting is impossible or not feasible within the time constraints common for most users. Some systems try to perform that task automatically, performing content-based filtering. They determine what articles should be rated as interesting or not by matching keywords or using rules that define the user's interests, in the form of a user profile [1]. There are different kinds of approaches to accomplish this. Some systems build the profiles automatically. However, sometimes human behavior is not easy to interpret [16] making this a difficult task. Other systems require direct intervention by the users to determinate their interests [7]. This causes undue cognitive loads resulting from the need to constantly classify the articles they read [13]. A third kind of system takes advantage of the usage patterns of several users to predict the current interests of another user [8]. It should not differ greatly from those displayed by others with similar profiles. However, not all users are equal to each other and this approach might not account for their individual variability [2]. A related problem occurs when a user's interests change dramatically in a short period of time. The system must be able to adapt accordingly.

To address those problems, we developed the WebClipping2 prototype. It allows people to access as many news sources as they want, from which it selects only the news articles that are more likely to interest a particular user in that specific moment. It uses the Bayes theorem to rate each document based on a profile built for each particular user [17]. This classification method is based on matching a set of keywords parsed from the news to be classified with another that defines the user's unique profile, representing its interests.

The user's interests are likely to change over time, and the system should be able to react to that change effectively. To maintain the users' profiles updated, we need some feedback from them to know how interesting they find the news they read. We could force them to classify every news article presented to them but that would be a task few users would be willing to do. To prevent that, we monitor the users' behaviors while reading

IUI'04, January 13–16, 2004, Madeira, Funchal, Portugal.

the news, to find out if they are interesting or not. Metrics like time, number of lines read and selected keywords are used to rate the article. Based on this passive feedback the user's profile can be continually updated and the next set of news will be more accurate rated. This method has proved to be very efficient in quickly adapting to radical changes of user's interests.

In the following section, we'll describe the overall structure of our prototype. Then, we'll discuss how the user profiles can be created and updated. How to infer the users' interests from the reading metrics will be discussed next. We'll then briefly describe some related systems, followed by some conclusions and considerations about possible future work.

2. THE WEBCLIPPING2 PROTOTYPE

The WebClipping2 prototype system consists of two distinct applications: one is responsible for the retrieval and filtering of news from the Internet; the other is a Palm news browser that also monitors the reading behaviors of the user.

The article filtering application accesses several news pages on the Internet and parses them, retaining only the relevant content and discarding all navigation, publicity, and other such page elements. It then rates the news articles it gathered according to the current user profile.

After the news are rated, they are sent to the PDA, where they can be read by the users. While doing so, several behaviors are monitored. The resulting metrics are fundamental for the construction and continuous adaptation of the user profile. This will facilitate the retrieval of more interesting news.

2.1 News Gathering and Parsing

The basis of all information handled by WebClipping2 are texts available on the Internet, either as written news articles, discussion forums, or any other informative documents. Most on-line information sources usually encode their content in HTML format, with all kinds of spurious elements not relevant to the content itself. Given the limitations of PDAs with regards to screen size and memory, it is important to filter the HTML pages to retain only the important information. Also, the news classification process will be made easier by considering only the news themselves. Fortunately, most news sites have a coherent, regular structure. The same elements (news titles, for instance) have always the same visual appearance, or are located in similar places. The now fairly common use of Cascading Style Sheets makes recognizing such elements easier since it specifically defines named styles to be used in a page. The HTML tags that are used are also sometimes identifiable. The fact that most news pages are automatically generated also helps keeping them self-consistent. This coherency allows us to establish a set of rules for the retrieval of news by parsing the contents of documents and isolating the relevant text.

In WebClipping2, it is possible at any moment to add, edit and delete news sites and to specify the necessary rules for their correct parsing. This requires a simple analysis of their content. To help users add new sites, we provide a graphical user interface where some templates can be used and parameterized. In fact, an informal survey of news sites showed that the majority of them belong to one of three templates in terms of organizational structure (Figure 1). These templates help users to identify the structure of a site, making the specification of the parsing rules easier. The application underwent several rounds of heuristic evaluation, allowing us do discover and correct several usability problems. After sources are parsed for individual news articles, these are classified according to their interest to the user and sent to the browser on the PDA.

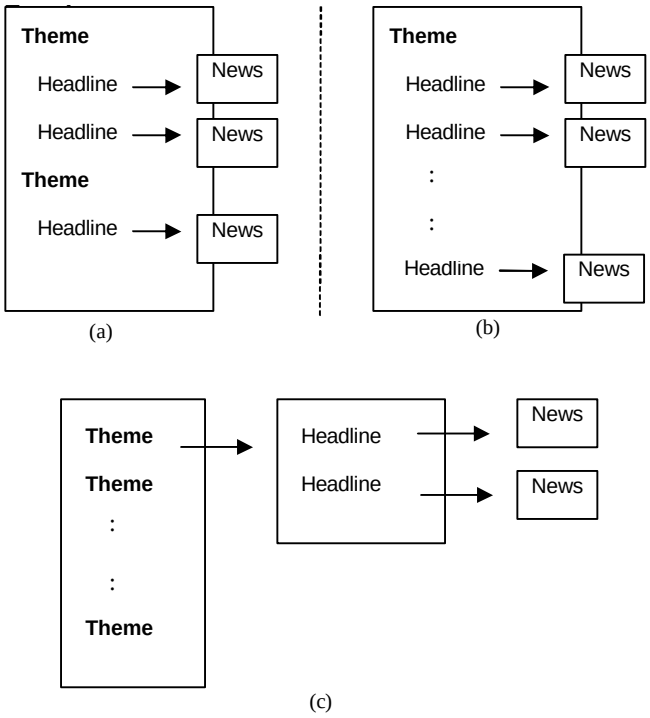

Figure 1 - Site Templates

2.2 Browsing the News

We implemented a special purpose news browser for PalmOS-based PDAs. This platform was chosen both due to its flexibility and popularity.

The browser's starting screen displays the news headlines, ordered by their interest level (the value with which they were classified). Once an article is selected, its contents are displayed. After the article has been read, and *only* if the user wants to, it can be rated according to how interesting it was. The user can also select specific words from the text as being of particular importance. Both actions are instrumental to updating the user's profile.

There are several limitations regarding the interface on the PDA. One of the most serious is the limited screen size that prevents the display of a large number of news headlines at the same time. To minimize this problem, we designed a *zoomable interface* that displays the headlines in several levels of detail according to their interest [6]. This alleviates the problem of reduced screen real estate and allows users to tailor the amount of news to be read according to the available time, ensuring the most interesting to be read first (Figure 2).

Figure 2 – PDA Browser, monitories user behavior

3. USER PROFILES

User profiles reflect the interests of users towards several subjects at one particular moment. They allow the system to classify the articles it fetched from the Internet according to their relevance for each particular user.

3.1 Creating the Initial Profile

We have a general database with up to 1,000 different words belonging to each of 31 different subjects. We created it by retrieving over ninety thousand news articles on those subjects from sites such as `ananova.com`. We then calculated the frequency of each word within each subject, and chose the most frequent to be part of our database. We took care of not including common words of the English vocabulary such as "he", "she", and "it". These were easy to eliminate automatically: they appear with similar probabilities in all subjects. While doing this, we discovered that certain words are frequent in more than one subject, not always with the same meaning. For instance, the word "space" is present in both "Astronomy" and "Travel". We took care of not eliminating words like these in borderline situations. We also noticed that some common words that apparently shouldn't be especially frequent in any one subject were present in the database. Later we came to realize that effectively, those words were relevant in specific contexts. For instance, words like "took" and "into" apparently don't have a relevant meaning by themselves, but in the "Sports" subject it is very usual for sentences such as "Agassi took Federer into second set" to occur.

This procedure has the advantage that it can be applied to any language, given a sample set of articles. Currently, we performed our tests in English due to an easy access to large numbers of on-line news sources in that language.

When they start using the system, users must specify their interests in some or all of the several available subjects, with a value ranging from 0 to 100. We then build a personal keyword database, as part of the user profile, by assigning to the words of each subject the value specified for it. This keyword database is vital to the news classification process

3.2 Keeping the Profile Up-To-Date

As the preferences of a user change with time, so will his profile need to be updated. The most direct way to know how interesting the users thought an article was requires them to explicitly classify it with an interest value on a scale of 1 to 4 (from 'not interesting' to 'very interesting'). Only four levels were considered to make the classification task as easy as possible. The article is then analyzed and the words therein that are part of the user's keyword database have their interest value updated in proportion to the classification (-10%, -5%, 5%, and 10%).

However, eliciting an explicit classification for all news articles a user reads is both intrusive and time-consuming. Classification tasks tend to cause undue cognitive loads and users tend to refrain from doing them. Furthermore, an explicit classification scheme would require awareness from the users that their interests are changing and there are no guarantees that the classification criteria will remain the same for all interactions with the system. Relying solely on the users' actions to keep their profiles up-to-date is not a good solution. Hence, we developed a way in which the user profile can be continuously updated automatically by the system, by taking into account how interesting the user found each specific news article and adjusting it accordingly.

4. INFERRING USER INTERESTS

To automatically infer how interesting a particular user found a news article, we developed a formula that allows the system to find that value based on the user's reading behavior. The browser collects, for each article, the following metrics:

- Total reading time, in seconds (**RT**).
- Total number of lines (**NL**).
- Number of lines read by the user (**NLR**).
- A constant (**k**), the user's average line reading time.

First, the percentage of the article that was read (**R=NLR/NL**) and the average reading time of one line for that particular article (**T=RT/NLR**) are calculated. The final inferred interest rate is given by **C=(T/k)*R**.

The only user-dependent aspect of this formula is the constant **k**. This constant is the average reading time of a text line for a user and can be determined by an initial calibration performed in the browser, where a sample "neutral" text must be read. This constant is necessary since people read at different speeds, and it is somehow necessary to be able to find if a particular article was quickly read or not (reading faster indicates a lower interest).

These formulas were developed iteratively resorting to user studies. We asked 23 persons to read and rate five news articles using a scale of 0 to 10, according to their interest. That classification was compared with the one inferred by the system. The formula was adjusted several times until it reached its present form. We verified that when users start to lose interest in an article, they quick scroll down to the end, or simply don't read the entire article. This might happen, however, not because the article is uninteresting, but because its essence was captured and

they feel no further reading is necessary. Our formula, by taking into account the percentage of the article that was read, correctly handles these two cases and appropriately classifies the article with a value similar to the explicit classification given by the user.

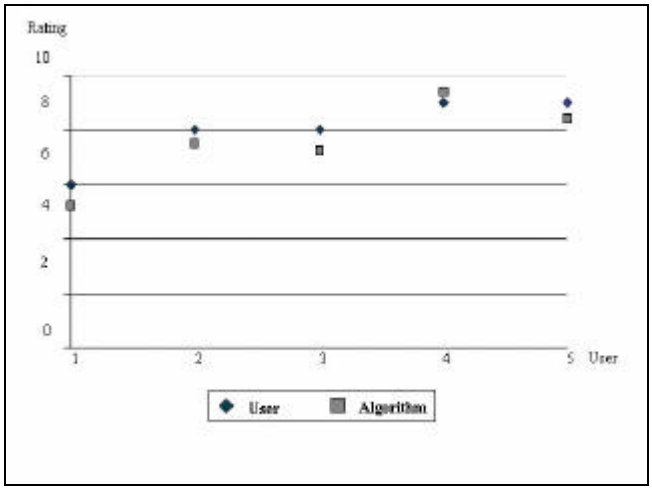

Figure 3 - Classifications by the user and the algorithm

We found a statistically significant correlation factor of 0.91 (95% confidence) between those values (Figure 3). In a normal usage situation, with only four interest levels, the accuracy increases even more.

5. CLASSIFYING THE NEWS ARTICLES

The classification of news articles is the core of our application, and is performed when they are retrieved from the Internet. Some content-based filtering is necessary so that only a small set of interesting articles is shown to the user, based on his profile. The classification associated to each article is an estimate of how interesting the user will find it. To compute that value, we use a Bayesian classifier. Given the conditional probabilities of finding a given word in an interesting article (present in the user profile), it uses Bayes' Rule to compute the overall interest of an article from the words therein. We use the following formula, the same used by Paul Graham on his email spam classifier [11]:

$$\frac{\prod p_i}{\prod p_i + \prod (1-p_i)} \quad ,i \in \{0 \dots 20\}$$

p_i – 10 best and worst classified words select by the parser

The reasoning behind this formula is straightforward. Each word has a value that reflects its probability of appearing on an interesting news article. Initially, those values are directly derived from the user's theme selections, and they are adjusted each time a user reads an article in which the words appear, according to his interest in that article. The classification of each article, given by the formula above, is nothing more than the combined probability of an article being interesting given the occurrence of its words.

After a tokenization process where individual words or tokens are identified, we compare them to the words in the user profile and apply the formula above. The resulting value is an estimate of the interest of the entire article. Articles whose classification is over a pre-determined threshold are shown to the user.

Since the user profile is created beforehand by the user from already existing sets of words, no training period is required. The initial results are sufficiently good to allow users to start using the system at once. The continuous usage will fine-tune the profile.

We consider only the ten best and worst classified words in an article, both for simplicity and efficiency's sake. Since the classifier depends only on the words that compose the continuously updated users' profiles, this algorithm is dynamic and able to perform a correct classification at any moment, reflecting the users' interests at that point.

6. RESULTS

To test the validity of the user profile update and news classification approaches, and the overall success of the application in terms of user satisfaction, we conducted several user studies. We considered three distinct user groups, corresponding to different initial profile configurations. We then monitored the behavior of the algorithms and evolution of the profiles over a two-week period, and conducted regular interviews with the subjects to estimate their satisfaction and the correctness of the process (if, indeed, they were presented with interesting news). The three distinct configurations were applied to eight different users, as follows:

- **Normal profile configuration:** The application was used without any restrictions. The users were free to define their initial profile (3 users).
- **Reverse profile configuration:** We asked the users to define their initial profile, but then reversed it (3 users).
- **Neutral profile configuration:** For these users, all selected subjects were given the same interest classification, half way across the scale (2 users).

6.1 Normal Profile

Since these users chose high values for their preferred themes, from a very early stage they received almost exclusively news that matched their preferences. After a while, some articles were classified with high values, between 0.9 and 1.0 (Figure 4). In fact, because the news the users read were, indeed, relevant, the inferred interests were high to begin with. When the already high values associated to the corresponding words in their profiles were increased, soon the top of the scale was reached.

We verified that the algorithm was capable of distinguishing interesting from non-interesting news that came from the same thematic site, even if they shared common subjects or traits: the latter obtained almost always an inferior classification. Some problems were found for User 2, as he received news classified as interesting that weren't (false positives) and vice-versa (false negatives). These mistakes occurred in an erratic way, and were caused by keywords belonging to more than one subject that had associated to it very high or very low values. When they occurred in a different context, they could negatively influence the

classification process. Despite these exceptions, the algorithm revealed itself to be able to filter out the really non-interesting news, which allowed all kinds of sites to be used as news sources, even if apparently not related to the most interesting subjects. If relevant news articles are published there, only those are retrieved. For instance, User 2 was interested in politics, but once got an article from a site of celebrity news because it mentioned Arnold Schwarzenegger's governor candidacy.

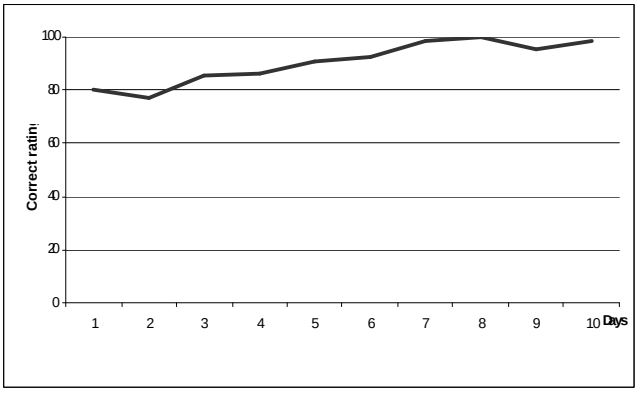

Figure 4 – Normal Profile Interest Rating Evolution

We also observed that there must be a distinction between the words explicitly selected by the user in the Palm browser from those already in his profile. Only that way can we ensure that the weights of the selected words are given a value proportional to those of other words and adequate for the user's interests.

6.2 Reverse Profile

The objective of this user group was to observe how the algorithm is able to monitor the user interests and adjust their profile accordingly. This test reflects the most extreme situation, in which the initial profile is exactly opposite of what it should be. The algorithm has to be able to completely reverse them to reflect the real interests. This was by far the most difficult test as the required changes to the profiles are radical. It was proposed as a challenge to the users, because a more committed and active collaboration was required to achieve the objective.

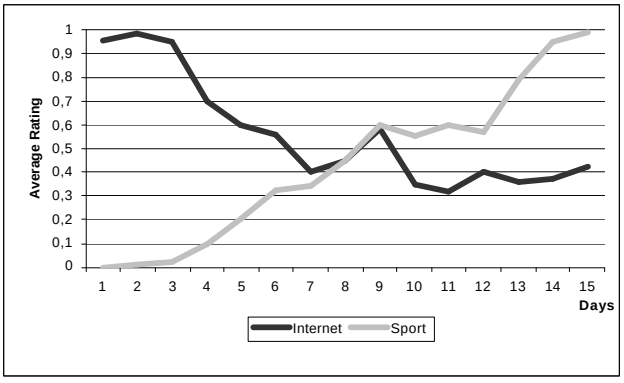

Figure 5 – Faster adaptation to user interest change

In an initial stage all the news were classified according to the reversed profile and so we had 100% of false-positives and false-negatives in the first few days. As time passed, and after some news were read (and others weren't), the values of the words in the users' profiles slowly started to reverse, due to the value oscillations induced by the interest displayed by the users in the several articles they were exposed to. Around the seventh to tenth day the quantity of false-positives and false-negatives suffered a notable reduction as the user profiles kept being updated. In the final days of the experiment the profiles were almost completely reversed, with only occasional badly classified news. The majority of articles were correctly classified in the terms of user satisfaction (Figure 5).

The reversal process happened in a regular and somewhat linear way. For User 6, the process was a bit slower because the subjects to be reversed were very similar (Business and Internet) to each other and some words belonged to both (Figure 6).

Through this configuration it was possible to verify that the algorithm can adapt in the case where users reveal a sudden interest in a new subject that was previously considered to be uninteresting. The complete profile reversal was accomplished within fifteen days of use of the application. The speed of the process seems to depend solely on the similarity between existing interesting and uninteresting themes (because of keywords belonging to both) and the users' behavior (how many news of each subject are read to determine which are considered interesting and non-interesting).

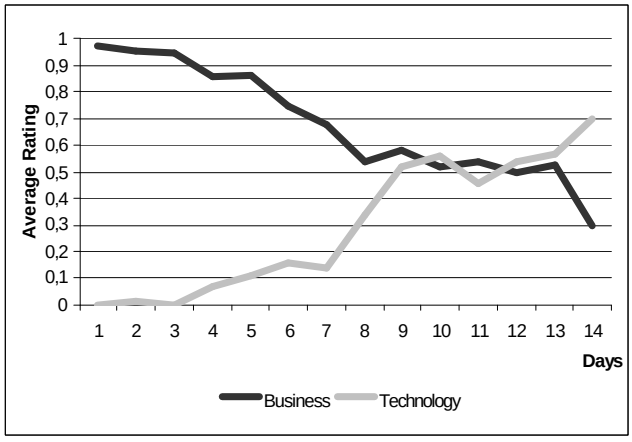

Figure 6 – Slower adaptation to user interest change

6.3 Neutral Profile

For this group of users the interest level was set as 0.5 for all themes they chose. The profile evolved, for User 7, as those of the users in the first group (Normal Profile), quickly adjusting itself to that user's interests. User 8, on the other hand, chose only one theme (Sports) stating that he was actually only interested in news of a particular sub-theme (Tennis). This allowed us to observe how the algorithm refines the values of the keywords in the users' profiles to reflect very specific interests.

In the first days, User 8 only read news relating to tennis, but the existence of words relevant to tennis but also to other sports resulted in unwanted false-positives and false-negatives. To help the profile change faster, the user explicitly classified some non-interesting articles as such. In the following days, the difference of the classifications of tennis and other sports related news grew larger, until after the fifth day almost all news were considered to be interesting (figure 7).

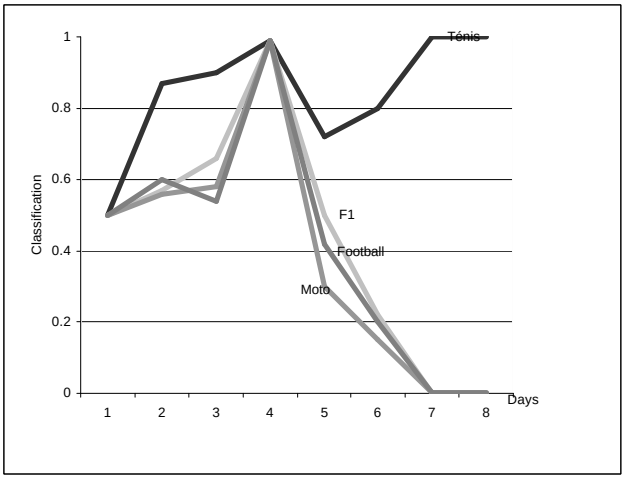

Figure 7 – Sub-Them Interest Evolution

Overall, we verified that, although some false-positives and false-negatives occur while the profile is still being defined or due to some sudden change of interests, some may still occur and are difficult to avoid. For example, User 7 had an interest on (real-life) Crime, but when a film about a murderer premiered the news concerning it (from a generalist site) were wrongly classified as interesting.

6.4 Discussion

All users stated their satisfaction with the system. Those that were allowed to use the system with their own profiles were satisfied after just one ore two days. Those for which the profile took longer to update were happy after eight to ten days.

The existence of inter-themes keywords may is one of the most disruptive factors in the classification process. If one of the themes is given a very high or low initial classification, its words will strongly influence the classification of the news articles. Thus, if one of those words appears on an article of an unrelated subject, it will tend to be classified in a wrong manner.

This occurs because the value of the words in the Bayesian classifier, when close to 0 or 1, produces overall ratings also near to 0 and 1. After some initial tests, we tried to minimize this effect by allowing only words explicitly selected by users in the browser to be classified with values greater than 0.9. In the definition of the initial profile, only values of up to 0.9 can be selected. This minimized the problems with inter-theme words.

The existence of a single word database in the profile implies that all words from all preferred themes the users choose are updated in the same manner independently of the subject they belong to. To separate the single database into one per subject would probably solve most of the problems with inter-theme words that could then be rated differently for each subject. However, this would require the system to know to what subject a particular news article belongs, in order to update the correct database. This would prevent the usage of generalist news sources and the ability to present news to the user that, while originating from usually uninteresting sources (as in the case of the Politics article in a Celebrity site, discussed above). Given those disadvantages, we chose not to follow this approach.

The size of the word database in the profile is also important, since if few of the words in an article are present in that database, it will be hard to produce a trustworthy classification. On the other hand, if the database is too large, most words will not be found in any one article, slowing the profile adapting process. A balance must be found.

Finally, while the values used to update the classification of keywords in proportion to the user's interest (-10%, -5%, +5%, +10%) are suited to quickly update the profile when there are interest changes, and cause no undue oscillation in quality when the profile is stable. However, such fixed values can make a word's quickly rise to extreme values that can result in classification problems. An adaptive approach, in which, the closer the value is to an extreme, the less it is changed, can minimize this behavior.

7. RELATED WORK

WebClipping2's main concern is to deliver to specific users news articles they consider interesting, in a mobile context, without the need for explicit classifications. Most existing approaches concern themselves with inferring interests on the World-Wide-Web. The Syskill&Webert system [14][15] requires users to explicitly state their interests. Letizia [12] and WebWatcher [5] suggest interesting pages and links, respectively. Other research projects such as those described in [9][10] and *Web Montage* [4] try to infer the interests automatically, resorting to web-specific metrics, such as the links in a page, mouse movement, etc. *Web Montage* requires a learning period to gather information about the users' navigation patterns and their preferences, such as the frequency visits to web sites visited, the time spent in each visit, and the links that were followed. Only after this training period is the system ready to offer suggestions to the users. Each time it is used, it generates a page with links to web sites that the user probably would like to visit at that moment, based on past behaviors. Webclipping2 requires no similar training period to discover the user's interests. The sources of information it handles are more uniform in nature than web sites in general. Allying this to a simple one-time interaction with the user where the initial interests are specified, WebClipping2 is ready to classify news articles, with a good performance, from the start.

None of these approaches correctly address the problem of inferring user interests in a mobile environment, where the interaction paradigm is different. On PDAs, no mouse is used. Time restrictions must be handled differently, and there often is no hypertextual structure to provide hints to related pages and their eventual interest. In [3], an approach that considers wireless devices and corresponding web navigation limitations is presented. It suggests links in real time, based on predictive models. In this approach it is possible to derive new sources of information (web sites) based solely on the user's behavior. However, unlike WebClipping2, only links to entire sites are presented, and no attempt to select the relevant pieces of information from a particular site is made.

8. CONCLUSIONS

Technology advances allowed information access to become easier. In fact, the Internet is now one of the favorite information sources, and its vast content is very appealing. However, it can often be difficult to find the right document in the right amount

of time. We showed how, in the context of mobile applications, it is possible to build and constantly update a user profile in order to represent the users' interests at any moment. We presented a solution that has no need for a training period, allowing users to take full advantage of the system from the start. Even in extreme interest change situations, the behavior of the algorithm is satisfying. In just ten days it can fully reverse a user profile to account for the user's real interests. The Bayesian classifier, while simple, proved to be an effective approach, although hindered by the occurrence of the same keywords in subjects with different interests. We also showed how reading metrics relevant for mobile-devices can be used to infer the users' interests in a simple but accurate manner.

9. FUTURE WORK

In order to take total advantage of the great amount of news available on the Internet, the parsing task should recognize the RSS format, increasingly used by news sources. This extension would remove complexity from the parsing process, since no spurious elements are present (RSS is an XML dialect). Also, more information related to the articles themselves is available, such as the authors, publication dates, themes and sources. This additional information would be important for a better classification performance.

Another relevant improvement could result from keeping the keyword databases separated for each subject, and classifying the news according to all of them, regardless of their source. Then, the highest value would be used. After the user has read that article, only the keywords on that particular database would be updated. This could help alleviate the classification problems due to inter-theme words without limiting the system's flexibility regarding news sources. Further studies are needed to verify this.

10. ACKNOWLEDGEMENTS

This work was funded in part by the Portuguese Foundation for Science and Technology, under grant POSI/34672/99.

11. REFERENCES

[1] Abbattista, F., Degemmis, M., Fanizzi, N., Licchelli, O., Lops, P., Semeraro, G., and Zambetta, F. Learning User Profiles for Content-Based Filtering in e-Commerce, *In Proceedings AI*AI Workshop su Apprendimento Automatico: Metodi e Applicazioni*. Sienna, Italy. 2002.

[2] Ahmad, M., Wasfi, A. Collecting User Access Patterns for Building User Profiles and Collaborative Filtering, *In Proceedings Proceedings of the 4th international conference on Intelligent user interfaces*, pp 57-64, ACM Press, 1998.

[3] Anderson, C., Domingos, P, and Weld, D. Adaptive Web Navigation for Wireless Devices**.** *In Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, pp 704-712, Seattle, USA. August 4-14, 2001.

[4] Anderson, C. and Horvitz, E. Web Montage: A Dynamic Personalized Start Page. *In Proceedings of the 11th World Wide Web Conference (WWW 2002).* ACM Press, Honolulu, USA. May 7-11 2002.

[5] Armstrong , R., Freitag, D., Joachims, T. and Mitchell, T. WebWatcher: A Learning Apprentice for the World Wide Web. *In Proceedings of the AAAI Spring Symposium on Information Gathering*, pp 6-12, 1997.

[6] Bederson, B., Hollan, J., Stewart, J., Rogers, D., Vick, D., A Zooming Web Browser, *Human Factors in Web Development*, Eds. Ratner, Grose, and Forsythe, Lawrence Erlbaum Assoc., pp 255-266, 1998.

[7] Billsus, D. & Pazzani, M. 1999. A Hybrid User Model for News Story Classification. *In Proceedings of the Seventh International Conference on User Modeling*. Banff, Canada, June 20-24, 1999.

[8] Bueno, D., Conejo, R. and David, A. METIOREW: An Objective Oriented Content Based and Collaborative Recommending System. *In Proceedings of the Twelfth ACM Conference on Hypertext and Hypermedia*, pp 310-314, ACM Press, 2001.

[9] Chan, P. Constructing Web User Profiles: A Non-invasive Learning Approach, *In Proceedings WEBKDD'99 Workshop on Web Usage Analysis and User Profiling*, pp 39-55, Springer-Verlag, 1999.

[10] Goecks, J., Shavlik, J. Learning Users' Interests by Unobtrusively Observing Their Normal Behavior. *In Proceedings Intelligent User Interfaces 2000*, pp 129-132, ACM Press, 2000.

[11] Graham, P. *"A plan for spam"*, 2002. available at http://www.paulgraham.com/spam.html.

[12] Liberman, H. Letizia: An Agent that Assists Web Browsing. *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pp 924-929, Morgan Kaufmann publishers Inc, Montreal, Canada. 1995.

[13] Ngu, D., Wu, X., SiteHelper: A Localized Agent that Helps Incremental Exploration of the World Wide Web, *In Proceedings of the Sixth International World Wide Web Conference (WWW6)*, pp 691-700. Santa Clara, California, USA, April 7-11, 1997.

[14] Pazzani, M., Muramatsu, J. and Billsus,. D. Syskill Webert: Identifying Interesting Web Sites. *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pp 54-61, Portland, 1996.

[15] Pazzani, M, Billsus, D., Learning and Revising User Profiles: The Identification of Interesting Web Sites, *In Machine Learning, 27*, pp 313-331, 1997.

[16] Schwab, I., Kobsa, A., Koychev, I., Learning About Users from Observation, *In Proceedings AAAI 2000 Spring Symposium: Adaptive User Interface*, pp 241-247, 2000.

[17] Versteegen, L. The Simple Bayesian Classifier as a Classification Algorithm, available on the Web at http://www.cs.kun.nl/nsccs/artikelen/leonv.ps.Z. 2000.