User Profile Generation Based on a Memory Retrieval Theory

Fabio Gasparetti and Alessandro Micarelli

Universita degli Studi Roma Tre,
Dipartimento di Informatica e Automazione
Via della Vasca Navale, 79 00146 Roma, Italy

Abstract. Several statistical approaches for user profiling have been proposed in order to recognize users' information needs during their interaction with information sources. Human memory processes in terms of learning and retrieval are with no doubt some of the fundamental elements that take part during this interaction, but actually only a few models for user profiling have been devised considering explicitly these processes. For this reason, a new approach for user modeling is proposed and evaluated. The grounds in the well-studied Search of Associative Memory (SAM) model provides a clear definition of the structure to store information and the processes of learning and retrieval. These assets are missing in other works based for example on simplified versions of semantic memory models and co-occurrence data.

1 Introduction

In a large repository of information, such as the Web, the importance of personalization is certain. The instant availability of many interesting resources is a great opportunity that users can exploit for many of their every day tasks. Nevertheless, when the amount of information exceeds the time a user spends to read and understand it, information overload issues have to be addressed. For this reason, the identification of the user information needs and the personalization of the human-computer interaction are becoming important research topics in this field.

There are a couple of important advantages if we choose to ground the personalization's user profiling on cognitive theories and models. Decades of studies in this field are available, each of them investigated through several kinds of evaluations. Furthermore, learning processes and memory structures could be considered as the foundation level where we can start building more sophisticated information behavior and seeking model instances.

Two systems use natural language processing and semantic or keyword networks in order to build long term user profiles and evaluate the relevance of text documents with respect to a profile: the SiteIF project [1] and the ifWeb prototype [2]. Even if that approach is sometimes called cognitive filtering, beside an implicit reference to simplified versions of the Quillian's semantic knowledge model, these two systems do not represent user needs with structures that are directly inspired by cognitive theories and models.

Gasparetti F. and Micarelli A. (2005).
User Profile Generation Based on a Memory Retrieval Theory.
In *Proceedings of the 1st International Workshop on Web Personalisation, Recommender Systems and Intelligent User Interfaces*, pages 59-68
DOI: 10.5220/0001422500590068

Two other works based on cognitive research but more focused on the prediction of user actions are: the SNIF-ACT model [3], and the Cognitive Walkthrough for the Web (CWW) [4].

In this work, we describe an approach to build user models that aims to emulate the process of learning and recovering of information that occurs in human memory. This is one of the innovative aspects compared with more traditional approaches, where the structure and the items represented are the major research subjects. We narrow the approach to the human-computer interaction in information environments where the user usually analyzes and selects documents in order to satisfy given information needs.

In the next section, we will briefly describe SAM, the general theory on which we have based our approach that is the subject of the following section along with all the advantages and possible issues. At the end, an instance of this model will be applied and evaluated in the context of the Internet browsing activities.

2 User Profiling

Before investigating in details the proposed user modeling approach, a brief introduction of the SAM theory is given. For a closer examination of this theory see for example [5].

2.1 SAM: Search of Associative Memory

The SAM is a general theory of retrieval from long-term memory that considers both the structure of the memory system and the processes operating within it. The structure refers to the items represented and their organization in the memory system, the processes refer to the main activities that occur within the memory, such as learning and retrieval.

The organization of the memory is divided in two parts: the Long-Term Store (LTS) and the Short-Term Store (STS). The STS shows two key features: it has a limited capacity and it is easily prone to forget its content. It can be seen as a temporarily activated subset of information enclosed in the permanent storage LTS. Its role corresponds to a working space for control processes, such as coding, rehearsal, decisions, etc.. When a new sensory input occurs, the related information is analyzed through the LTS structure. The result of this analysis is the activations of some information units that are placed in the STS. It is also possible to update the STS by means of internally generated probe cues, composed of information previously retrieved from the LTS and currently still in the STS.

Both kinds of memories are composed of unitized images, the objects that may by sampled and recovered during a memory search. The retrieval process is based on the associative relationships between probe cues and memory images that can be represented by means of a matrix of retrieval strengths. These strengths determine the probability for a given set of probe cues to gather an image stored in the LTS. Basically, the user model is built by means of all the strengths stored in that matrix.

In the retrieval process, along with the current cues, it is also possible to consider the contextual information related to the current user's task, such as, temporal data,

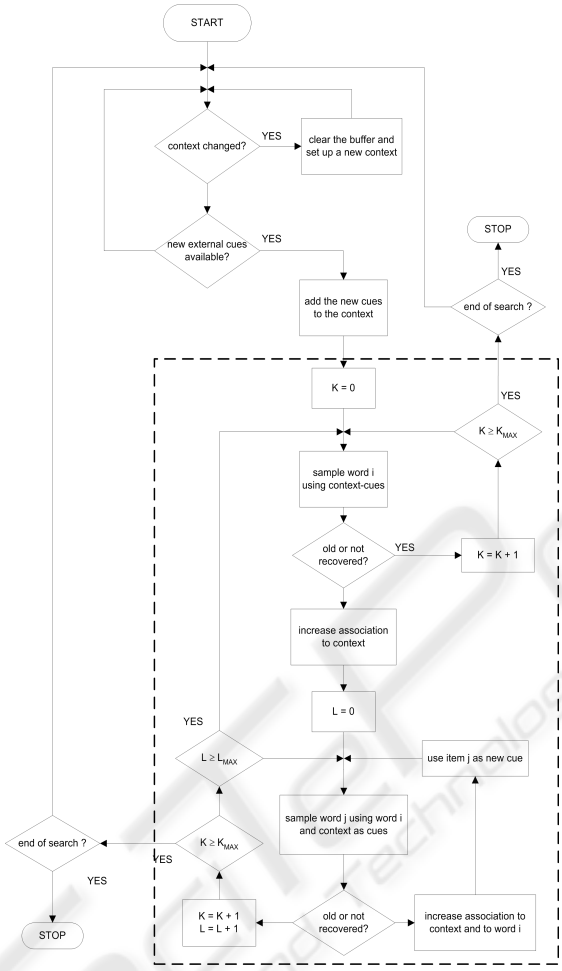

Fig. 1. An extension of the retrieval process flowchart in Raaijmakers and Shiffrin 1981 to consider the external cues originated by the interaction of the user with the information sources and theirs effect on the process. The broken box margins the original chart.

category and item names, etc.. Actually, the current cues used during the learning and retrieval can be seen as a context as well. If we store an image associated with the current cues, some relationships between this image and the cues are created, or increased if they already exist. Each time we have the same cues as context, such an image will tend to be retrieved. As we will describe in the next section, the context-image relationship is viewed as an essential feature in the user modeling approach.

The LTS is probed with the current information related to the user's task, the information available in the STS and the one retrieved earlier in the search. This process determines what image is sampled and made available to the user for evaluation and decision-making. It is convenient to separate the sampling phase in two parts: (1) Be-

cause just a small number of images in the LTS have a non negligible strength to the current probe cues, an initial set is built including those images in LTS, (2) After the sampling phase, the image elements that will be activated and made available to the user is drawn. This recovery process depends on the strength between the selected images and the probe cues as well.

For the building and updating of the LTS structure, it is possible to use the same storage approach followed in the SAM Simulation (SAMS). This approach consists of a buffer rehearsal process [6] that updates the item-item and context-item strengths as a function of the total amount of time the pair's objects are simultaneously present in the rehearsal buffer. This buffer corresponds to the STS, so it has a limited size and stores all the cues plus the current context. If the buffer size is reached, an item will be randomly replaced.

The computational details of the theory, along with all the adaptations we made for the context under examination will be discussed in the next section.

As the authors of the SAM theory have emphasized, the long-term memory images and probe cues are quite distinct. The images in memory correspond with the past cues plus the context at the moment of the learning. This relationship is used in the sampling process to identify the images related to a given context, but also to be able to retrieve all the connected cues stored during the learning.

2.2 The Proposed User Modeling Approach

During each information seeking session where the user is looking for sources able to satisfy particular information needs, the user must assemble and deal with a set of concepts related to these needs. If we were able to identify all these concepts it would be possible to personalize the interaction with the sources and, at the same time, filtering the information that could be retrieved by autonomous search tools. Because human memory is involved each time the user runs into a new interesting concept, whether he has to learn or retrieve it from his memory, a technique that aims to emulate this memory could be successfully used as user modeling component in a personalized system.

Unfortunately, the user does not usually confine the interaction to one or few information sources. Talking to colleagues, friends and relatives, reading newspapers and magazines, watching TV are just examples of the possible interactions a user can be involved in. Nevertheless, we assume that given a information need, if we look at the different kinds of interactions, the user will not ignore the most important concepts. Therefore, if we restrict our analysis to one of the possible interactions, it is always possible to recognize the subset of the most valuable concepts.

Although the SAM theory does not require a particular memory representation, for the task under consideration we need to define how the information is encoded. In this study we have chosen the word as the atomic unit of information that is possible to store in the LTS and STS. This choice is primarily motivated by its simplicity and adaptation to the personalization context. We are considering environments where the information is represented by text sentences, so the word seems the obvious element that could be stored and analyzed.

Natural language processing techniques could be also used to raise the representation level of the units. If we were able to identify the concept related to a particular word

or sentence, we could actually store this information in the memory structure, getting closer to a semantic network definition. However, due to the low reliability of these techniques, in this study we have decided to restrict the representation to the word.

As we have explained in the previous section, the retrieval process consists of two steps: the sampling of the images in the search-set, and the recovery of the information contained in the sampled image. Both steps make use of a *strength matrix*, where the generic value $S_T(Q_{ij}, Ij)$ corresponds to the strength between a cue Q_{ij} and the image I_j. The rows Q_{ij} refer to the images, which in our case are words, I_j and the Q_{i0} values refer to the context cue.

The first step draws all the sampling probabilities of sampling a given image I_i in the LTS as a function of the current cues Qs:

$$P_S(I_i|Q_1, Q_2 \dots Q_M) = \frac{\prod_{j=1}^{M} S_T(Q_j, I_i)^{W_j}}{\sum_{k=1}^{N} \prod_{j=1}^{M} S_T(Q_j, I_k)^{W_j}} \quad (1)$$

where M is the number of cues, and N is the matrix dimension. The W_j values are weights used to set the importance of the single cues. In our domain for example, these weights could be set as a function of the Inverse Document Frequency (IDF) values in order to decrease the weights of the words that frequently occur in a given corpus and that can be considered very common and therefore little relevant.

This formula gives the highest probability to the terms with the highest product of strengths, and hence those that tend to be greatly associated to all the current cues.

Once a word has been sampled, the recovery process step takes place. For the sampled image I_i we draw the probability:

$$P_R(I_i|Q_1, Q_2 \dots Q_M) = \frac{\sum_{j=1}^{M} S_T(Q_j, I_i)^{W_j}}{\sum_{k=1}^{N} \sum_{j=1}^{M} S_T(Q_j, I_k)^{W_j}} \quad (2)$$

The currently activated information in the STS and its relationship with the one in the LTS is used to build new relations and structures in the memory. The important feature of the theory kept in our approach is that the information currently active in the STS tends to be stored together. So in each storage phase, an item is stored along with all the contextual information, the same information that could help retrieving the item later, even if the STS misses a direct reference to it.

In natural language domain, such as the one under consideration, the context helps also to disambiguate the meaning of a word stored in the LTS. The relationships among items bind each word with the context at the time of learning. If the user runs into the same word but with different contexts, different sets of relationships will be created between the item and the contexts. During the retrieval, if we probe just the ambiguous item it is not possible to have any information that helps us to recognize its meaning. But if we add just one element contained in one of the contexts, the retrieval process has much more chance to return other items related to the context at issue.

Moreover, the importance of one term for the user is not drawn just by a single value or weight associated to it, as it happens in the Vector space model or in many Bayesian approaches. A term is judged interesting if it is also bound to a context that is related to

the current context. In other words, the importance is not longer an becomes an absolute value but a relative one.

The learning process that corresponds to the process described in the SAM theory updates the strengths in the LTS as a function of the total time a word or a pair of words is stored together in the STS buffer. Given t_i the time spent in the buffer by the generic image, i.e., word, I_i, and given t_{ij} the time the images I_i and I_j occur together in the buffer, we have:

$S_T(C, I_i) = at_i$

$S_T(I_i, I_j) = S_T(I_j, I_i) = bt_{ij}, \;\; t_{ij} \neq 0$

$S_T(I_i, I_i) = ct_i$

If a word pair have never appeared together in the buffer, they assume anyway a non negligible residual retrieval strength d. The values a, b and c are parameters of the model, and C is the context currently in the STS.

The strengths are also updated each time a particular combination of cues is able to sample and recovery a particular image. In this case the strength between the cues and the sampled image, and the self-association strength is incremented:

$S'_T(C, I_i) = S_T(C, I_i) + e$

$S'_T(I_i, I_j) = S'_T(I_j, I_i) = S'_T(I_j, I_i) + f, \;\; t_{ij} \neq 0$

$S'_T(I_i, I_i) = S_T(I_i, I_i) + g$

where the S are the strengths before the incrementing and e, f and g are other parameters.

Now that we have explained how the user model is built and updated, the general retrieval process in Fig.1 that occurs each time the user interacts with a information source is described. This process aims to simulate as much as possible the analysis and recovery of the concepts that the user goes through during the information seeking process.

The process is composed of the inner retrieval process that resumes the one discussed in the SAM theory. In the first cycle there are at the most K_{MAX} attempts to sample and recovery a word stored in the LTS by means of the current context. A failure is every attempt that does not lead to recall of a new item. This may happen either because of low recall probabilities, Eq. 2, or if the image has been already unsuccessful sampled in a previous cycle and the context has not been altered meanwhile.

If a word has been successfully sampled, the strengths to the current context $S_T(C, I_i)$ and the self-association $S_T(I_i, I_i)$ are incremented, and the word is added to the context which is used in the next cycle to sample other correlated images. If the recovery fails, than the counter K and L are incremented to see if we are at end of the cycle. Each time a word is sampled in the second cycle, it joins the context and the strengths $S_T(C, I_i)$, $S_T(I_j, I_j)$ and $S_T(I_i, I_j)$ are updated.

The outer elements of the flowchart concern the information that comes from the interaction between the user and the information sources. Here we consider each data that has some kind of relation with the user's information needs and that can be represented in a text format. Possible examples are: queries, document's snippets or categories' keywords selected by the user.

When new information is available it is added to the context and a new retrieval process takes place to recovery all the correlated information and, at the same time,

store the data in the LTS. At the end of the retrieval process, we check if the user has completed his seeking session, in this case we leave the main cycle of the process. If the user is still trying to satisfy his information needs, the old context is compared to the current one in case the user has changed the domain in which he is working on. A new context means a reset of the STS, that is, all the temporary information stored during the session up to the context change is wiped out.

Each time information is available and the recovery process is completed, we are able to collect the information stored in the STS that can be used to give a representation of the current information needs of the user. This representation regards the last context that has been identified during the search.

It is possible to see how the approach can work without the user's explicit feedback, especially if the cues originated from the user interaction with the information sources are fairly representative. Beside taking time during the seeking processes, past studies showed how explicit feedbacks are not able to considerably improve the user model especially if a good interface to manage the model is not provided [7].

A last remark on this approach is the absence of the *renting* technique often used in order to remove concepts from the user model that are no longer judged interesting for the user. There are two reasons to justify the presence of this technique. During the information behavior the user learns concepts that influence his knowledge and therefore his representation of the information needs. A user model should be able to recognize these knowledge alterations and their influences on the user's goal. The second reason regards the accuracy of user models. Many content-based techniques take a representation of the documents' content seen by the user as the primary source to build the model. But many times documents concern with different topics, and the user is interested in just a subset of them. If we update the model with irrelevant information, a technique to *forget* some concepts no longer considered is needed in order to improve the model as time goes by.

One of the interesting features of the discussed approach is that the model is affected just by an additive learning process. It does not mean that it is impossible to ignore a concept that is judged not interesting. Even if no deletions can occur, the learning process tends to increase the strengths of the relationships of concepts frequently covered during the seeking process. A wrong concept occurred during the process is stored in the LTS but its strengths with other images will not be intensified in the future. As a result, the probability to recovery this item given a context will get low values. In other words, forgetting is a result of the failure in the attempt to retrieve a concept.

2.3 An User Model Instance for Browsing Activities

In this section, we want to briefly describe an instance of the proposed user profiling approach in a concrete domain, as the Internet browsing. This is a very important domain because of the amount of the available data that can be exploited during any information seeking task. That is one of the reason the personalization of the interaction with the search tools is becoming an important research topic. Identifying what a user is currently looking for during a browsing session, that is his information needs, is the first step toward an efficient personalization.

In order to have an instance of the user modeling approach, a methodology to identify the cues that will be used in the learning and storage process discussed in the previous section is needed. The notion of Information scent [8, 9] developed in the context of Information foraging theory [10], is a valid choice to identify these cues by means of the text snippets associated to the links. Users use these browsing proximal cues to decide if access to the distal content, that is, the page at the other end of the link. Formally, the information Scent is the imperfect, subjective perception of the value or cost of the information sources obtained from proximal cues.

Therefore, it is possible to begin a learning and retrieval process each time a user selects a link and visit a page, considering the anchor texts of the chosen link as the cue. But from preliminary evaluations we have verified that this information could be not enough to recognize valuable cues for the learning process, especially if the text consists just of few words with no correlation, e.g., “full story”, “page two”, “link”, “rights reserved”, etc..

For this reason we have developed an algorithm whose goal is to collect all the information related to a given link selection. In a few words, the text of a link is joined with the title of the linked page that has been visited by the user. Then the context of a link, i.e., the text surrounding the anchor is retrieved by means of the page’s Document Object Model tree representation. In other words, the page is divided in units whose boundaries are arranged by a subset of HTML tags, e.g., TR, P, UL, etc., and the text of the deepest unit that contains the link is set as the context of the link. At the end, this context is compared to all the other units of the page that includes the link, in order to find further related text for the cue. The comparison is based on the same IR similarity function used in the TextTiling algorithm [11].

For simplicity, in this evaluation we have ignored the explicit representation of contexts. Actually a context is an important facet of the approach because it let the model identify multiple topics and create for each topic particular relationships with the images, i.e., words, unlike many other modeling techniques where each word is evaluated as important or not regardless of the current context. But it is however possible to obtain interesting results taking under consideration just the implicitly relationships between a word and the context represented with the other words currently stored in the STS. This alteration affects the retrieval process replacing the $S_T(C, I_i)$ with the set of $S_T(I_j, I_i)$ for all the I_j images currently in buffer.

3 Evaluation

In order to evaluate the proposed user modeling approach, we have collected a set of browser histories from a certain number of users. After a filtering process needed to remove not interesting Web sites, e.g., Web mail services, we have selected browsing subsequences related to topics that we were able to easily recognize. An external person selected a subset of these subsequences and, therefore, a subset of the topics. The resulting subsequence set has been the input of our user model.

A traditional user model has been employed to provide a benchmark to compare our results. It is based on a Vector Space model (VSM). A Relevance Feedback (RF) technique is used to update the model with the information collected during the browsing.

The input of the two user models is the information collected by the technique described in the previous section based on the anchor texts, the titles and the correlated text in a Web page. The images stored in the user model correspond to the word extracted from this information. A more refined approach, where an image is associated with the meaning of the word has not yet been considered.

During the evaluation, we have submitted the selected subsequences. At the end we have collected 25 keywords with the highest score from the VSM-based user model, i.e., the highest frequency of occurrence. For the proposed user model, we have instantiated the recovery process described in Sec. 2.2, see Eq. 2. In this way we are able to collect the same number of keywords related to the given probe cue, as a function of the information stored in the model.

An external judge assigned a score in 4 point relevance Likert scale for each keyword extracted from the user models, from no-relevance to high relevant, according to the topic under consideration.

The parameters of the model get these values: $a = 0.1$, $b = 0.1$, $c = 0.1$, $d = 0.2$, $e = 0.7$, $f = 0.7$ and $g = 0.7$. The maximum dimension of the rehearsal buffer is 100 elements.

Table 1. Evaluation results for the seven sequences of browsing histories analyzed building a single user model.

Topic	VSM + RF	Proposed Approach w/Input
New England Execution	0.16	0.8
Green Day Band	0.2	0.16
Podcaster	0	0.4
Tsunami	0	0
Ray Charles	0.24	0.76
Popular T-Shirt	0.16	0.8
Vegetarian Diet	0.16	0.2

The results are shown in Table 1. It is possible to note how the proposed user model outperforms the traditional VSM approach. This phenomenon is related to inability of the VSM-based user model to store browsing contexts. The evaluation shows how the instantiation of the proposed user modeling approach is able to recognize important keywords as a function of the current context, i.e., the probe cues, in comparison with a traditional user model based on the Vector Space model and the Relevance Feedback technique. Once we are able to identify those keywords, it is possible to use that model for example in the filtering task, re-ranking the results of a search engine by means of the keywords returned from the user model after a browsing session, or building hypertext pages adapted to the user needs.

4 Conclusion

Beside describing an approach to build user profiles, which aim to recognize concepts related to the current information needs, this work suggests a standpoint where the user profiles are seen consisting of multiple layers. The low level regards the basic human memory processes of learning and retrieval while the higher levels deal with the information seeking strategies a user can undertake in order to reach a particular task.

An implementation of a user model based on the proposed approach has been briefly showed along with an evaluation that has produced interesting results. The future work will concern the extension of the approach to include other information that can be collected, such as the previous contexts occurred during the learning or the category of the current images. The latter information can be easily obtained by one of the common categorization techniques in the literature.

References

1. Magnini, B., Strapparava, C.: User modelling for news web sites with word sense based techniques. User Modeling and User-Adapted Interaction **14** (2004) 239–257
2. Asnicar, F.A., Tasso, C.: ifweb: a prototype of user model-based intelligent agent for document filtering and navigation in the world wide web. In: Proceedings of Workshop Adaptive Systems and User Modeling on the World Wide Web (UM97), Sardinia, Italy (1997) 3–12
3. Pirolli, P., Fu, W.T.: Snif-act: A model of information foraging on the world wide web. In: Proceedings of the 9th International Conference on User Modeling, Johnstown, PA, USA (2003)
4. Blackmon, M.H., Polson, P.G., Kitajima, M., Lewis, C.: Cognitive walkthrough for the web. In: Proceedings of the ACM conference on human factors in computing systems (CHI2002), Minneapolis, Minnesota, USA (2003) 463–470
5. Raaijmakers, J.G., Shiffrin, R.M.: Search of associative memory. Psychological Review **88** (1981) 93–134
6. Atkinson, R.C., Shiffrin, R.M.: Human memory: A proposed system and its control processes. In Spence, K., Spence, J., eds.: The psychology of learning and motivation: Advances in research and theory. Volume 2. Academic Press, New York, USA (1968) 89–195
7. Wærn, A.: User involvement in automatic filtering: An experimental study. User Modeling and User-Adapted Interaction **14** (2004) 201–237
8. Chi, E.H., Pirolli, P., Pitkow, J.: The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI2000), Hague, Netherlands (2000) 161–168
9. Chi, E.H., Pirolli, P., Chen, K., Pitkow, J.: Using information scent to model user information needs and actions on the web. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI2001), Seattle, WA, USA (2001) 490–497
10. Pirolli, P., Card, S.K.: Information foraging. Psychological Review **106** (1999) 643–675
11. Hearst, M.A.: Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics **23** (1997) 33–64