Introduction to Part II

In Part I, we studied how the context of a stimulus influences the retrieval processes that are triggered by that stimulus. For this purpose, we studied small-scale behavioral effects and provided explanations for the observed effects by extending the declarative memory system of the ACT-R architecture of cognition. The most important conclusion that can be drawn from Part I is that memory retrieval is not a static, ballistic process, but rather a dynamic process that is adaptive to current environmental demands.

In Part II of this thesis, we will apply this insight in a larger context, and study how prior knowledge and the current environment interact in predicting which concepts will be retrieved from memory. The cognitive models that we will develop in this Part, will be used for recommending relevant items, first in the cultural heritage domain (Chapter 5), and then in the domain of scientific literature search (Chapter 7). This way, we will explore whether these so-called recommender systems may be based on principles from cognitive science. That is, the main concern in Part II of this thesis is whether in principle it is possible to develop useful recommender systems that incorporate models of memory. Part I of this thesis will act as the theoretical basis of these models, in the sense that many concepts developed or discussed in Part I will be applied in the context of cognition-based recommender systems in Part II.

Chapter 5 will discuss how a memory model can be used to develop an artwork recommender system that selects artworks for presentation that are interesting to an individual museum visitor (the Virtual Museum Guide or VMG). The VMG combines the perceived interests from its users with its knowledge on the museum’s collection to provide a personalized museum tour through the online collection of the Rijksmuseum Amsterdam (www.rijksmuseum.nl).

Chapter 5 will also discuss the results of a user study that we performed with the VMG. In this study, the visitors that were presented a tour constructed by the VMG did not rate the tour as better suited to their interests as visitors who saw non-individualized tours. This unexpected result may be due to two assumptions that we made in developing the VMG. Chapters 6 and 7 report work that assesses the effects of these assumptions.

The first assumption underlying the work in Chapter 5 is that binary feedback (“Interesting” vs. “Not interesting”) given to the artworks presented by the VMG is sufficiently detailed to develop an accurate user model of the visitor that reflects his or her art interest. Given that no effect of personalization was found, it might be that this measure was too coarse. In Chapter 6 we tested whether eye gaze is a better measure of interest in artworks than binary feedback. In Chapter 6, we will present a study in the cultural heritage domain, in which the visitor’s eye gaze is used to determine which information on the observed artwork is presented next.

The second assumption we implicitly made in developing the VMG, is that the interest of museum visitors is mainly related to the cultural-historical value of the artwork. This was at least partly due to the annotations of the Rijksmuseum that focus mainly on the cultural-historical values. As we based our recommendations on these annotations, we modelled the VMG’s knowledge on the museum’s collection as a semantic network (Collins & Loftus, 1975; Quillian, 1968) of cultural-historic and art-related concepts. Thus, we implicitly assumed that visitors would have a similar interest in artworks that had similar cultural-historical values. This means that we ignored other aspects of the artworks that also may have influenced visitor interest. For example, the use of certain colors, painting or crafting techniques, or a particular...
arrangement of figures in the scene may also contribute to the appeal a work of art may have to a visitor. These aspects relate to the expressiveness of an artwork (Arnheim, 1954/1974), which is not easily described. For this reason, we turned to the domain of scientific literature search (Chapter 7), in which the relation between interest and semantic similarity is clearer. That is, when searching for scientific literature, it is highly likely that a scientist is interested in those papers that share certain keywords with the scientist's own published work.

In Chapter 7, we will discuss a recommender for scientific papers that is based on formal models of declarative memory, the Personal Publication Assistant (Publication PA). The chapter discusses how the history of usage of words in published abstracts of researchers can predict which abstracts an individual researcher currently finds of particular interest. We will present an experiment that shows that participants preferred abstracts that were selected for them by the Publication PA to abstracts that the Publication PA deemed uninteresting for them.

In Chapter 8, we will study the performance of a memory-based recommender system for scientific literature search, by comparing it to a set of often-used decision-making algorithms. The reason for this comparison is that, although we refer to scientific literature search as a problem of information selection, it can also be perceived as a problem in the broader class of decision-making problems. That is, for every abstract, the recommender system has to decide whether to recommend it. To compare our system with a broad array of alternative techniques, we chose different types of alternative algorithms. Some of the algorithms in the comparison are chosen because they provide a performance benchmark in the decision-making literature (e.g., take-the-best or multiple regression, Gigerenzer, Todd, & the ABC Group, 1999). Others are included because they share features with the Publication PA. The results of Chapter 8 show that the Publication PA outperforms the other competitors that were trained on the same data set for some users, but is outperformed for other users.

Based on the results obtained in Chapters 7 and 8, it seems that it is possible to develop a useful recommender system based on a memory model. However, it is important to consider the domain for which the recommender system is being developed. In particular, it is important to assess whether the cues that are going to be stored in the declarative memory model represent the relevance of the items that may be selected. In the domain of literature search, this is the case, resulting in a positive evaluation of the memory model. However, in the cultural heritage domain, this seems not to be the case (Chapter 5). Here, the recommendations of the VMG did not align with the interests of the museum visitors, possibly because the cues, cultural-historic concepts, did not represent what museum visitors find important when perceiving a work of art.
Artwork selection for a Virtual Museum Guide

INTRODUCTION

With the advent of online information presentation, cultural heritage institutions are starting to make their collections available online. Many museums already have websites displaying digital reproductions of parts of their collection. Some of these online repositories are annotated, making it possible to search for specific artworks: For example, the website of the Amsterdam Rijksmuseum in The Netherlands is driven by an ontology on art and artists.

With the online presentation of cultural heritage content, new issues arise. While one of the advantages of digitalization and online presentation is the greater accessibility of cultural heritage (e.g., because of better search capabilities, Van Ossenbruggen, Amin, Hardman, Hildebrand, Van Assem, Omelayenko, Schreiber, Tordai, De Boer, Wielinga, Wielemaker, De Niet, Taekema, Van Orsouw, & Teesing, 2007), one of the drawbacks is that there is less control over what is presented to an individual visitor. Cultural heritage institutions have as one of their aims to educate people on history and culture, which becomes harder to realize once the contents of their collection is accessible from anywhere; They can no longer cater the individual interests of museum visitors while maintaining coherence in the presented information. Besides the decreased control that cultural heritage institutions experience, finding interesting artworks in an online museum poses a problem. Just like in a real museum, most online museum visitors are not aware of their specific interests or of the exact contents of the museum’s collection (Bell, 2002). Instead, they only have a general impression of what they want to see and what is available. This makes it difficult to adjust the presentation of the artworks to the visitors’ personal interests.

Consider the example of a professional, educated museum guide, touring a party of interested visitors through a museum. The guide can (and has to) select information on the artworks from her extensive knowledge that relates to the personal interests of the party, and can choose which artwork to present next from the collection on display. To reproduce a similar personal experience in an online setting, personal interests as well as relationships between artworks have to be known. A successful recommender system for the cultural heritage domain should incorporate both issues mentioned above: On the one hand, it should take care of the educational role of a cultural heritage institution, and on the other hand it should provide an enjoyable and personalized experience.

OVERVIEW

In this chapter we will present an online recommender system that presents artworks from the Amsterdam Rijksmuseum collection. In our approach we will try to model the way a human museum guide will behave while touring a visitor through a museum. In order to achieve this, we will ground the structure of the recommender system in cognitive theories (Anderson, 2007a).

To stress the analogy with a museum guide touring a group of visitors through a museum, we termed the system the Virtual Museum Guide (VMG). The VMG combines the relationships

This chapter is an extended version of Van Maanen, L. (2007). Mediating expert knowledge and visitor interest in artwork recommendation, IWA 2007: Halle (Saale), Germany.
that artworks have to each other with the personal interests of the visitor to arrive at suitable art recommendations. We will first give an overview of the most important aspects of the system, and then discuss each aspect in more detail.

In the system we will present here, the artworks presented online are accompanied by sets of keywords that describe the interesting aspects of the artwork. As these keywords are provided by the museum’s art experts, expert knowledge on the artworks and their interrelations are contained therein. We have applied statistical inference tools from natural language research (Landauer, Foltz, & Laham, 1998) to infer how the artworks relate to each other (details will be provided in the section on the Knowledge Base below). This way, all artworks are related to each other with an association value indicating the relevance of one artwork for another. This structure can be thought of as a semantic network or spreading activation network (Collins & Loftus, 1975; see also Niessen, Van Maanen, & Andringa, 2008; Quillian, 1968).

For the system presented here we opted for the use of an explicit interest indicator using an Interesting and a Not interesting button. Each time a user indicates interest in an artwork by clicking one of the two interest-buttons, this feedback is stored as a declarative chunk in the VMG’s memory. All chunks representing a visitor’s feedback form a user profile of that particular visitor (the Visitor Model). A new artwork will be selected by computing the most relevant and interesting artwork, given the current state of the Visitor Model. First, the visitor’s interest in the already visited artworks will be assessed. Second, a spreading activation algorithm (described in more detail below) is used to compute a combined measure of interest and relevance.

THE RATIONAL ANALYSIS OF MEMORY

The Virtual Museum Guide’s memory is based on a formal theory of human declarative memory, referred to as a rational analysis of memory (Anderson & Schooler, 1991), which we will apply for the development of a recommender system. The key idea of the rational analysis of memory is that human memory is optimally adapted to deal with information that has been presented in the past (Anderson & Milson, 1989; Anderson & Schooler, 1991). Anderson and Schooler (1991) demonstrated that for each declarative fact stored in memory, the probability that that piece of information will be needed in the immediate future reflects the history of usage of that piece of information. That is, information that has been presented recently is more likely to be needed again than items that have been presented in the more distant past. Also, information that has been presented more frequently is more likely to be needed again. Often, the probability that information will be needed in the immediate future is represented by a quantity called activation. The declarative memory representation consists of small pieces of declarative knowledge, called chunks, which together represent a person’s long-term memory.

The two environmental observations (recency and frequency) have crystallized (Anderson et al., 2004) into the following equation:

\[ B_i = \ln \left( \sum_{j=1}^{n} \frac{1}{\sqrt{t_j}} \right) \]  

\[ \text{equation 5.1} \]

\( B_i \) represents the base-level activation of a chunk (indicated by the index \( i \)). Equation 5.1 captures the effect of frequency of presentation by summing over multiple presentations, and the effect of recency of presentation by dividing by the square root of each presentation time lag (represented by \( t_j \)), that is, the time since each presentation of a chunk. This equation
has been used in numerous studies predicting memory retrieval effects, both for theoretical purposes (e.g., Anderson et al., 1998; Van Maanen & Van Rijn, 2007b) and for application-based research (e.g., Pirolli, 2005; Van Maanen et al., in press).

Besides the frequency and the recency with which chunks are encountered, also the contexts in which they are encountered add to their activation. The context effects between chunks are represented by an association value that reflects the likelihood that two chunks have co-occurred in the past (Anderson & Milson, 1989). The association between chunks predicts how relevant each chunk is in the context of another chunk.

We used a rational analysis of memory to develop a model of a museum guide. The VMG will remember (and forget) the interests of the visitor in a way that is similar to human behavior. That is, the VMG will compute which artworks are relevant in the context of its recollection of a visitor’s previous feedback. We incorporated the association between artworks by computing the semantic similarity between two artworks, based on the likelihood of co-occurrence of certain keywords that relate to the artworks.

VIRTUAL MUSEUM GUIDE

Figure 5.1 presents a flowchart of the Virtual Museum Guide. We start out with extraction of a Resource Description Framework (RDF) specification of each artwork from the online ARIA (Amsterdam Rijksmuseum InterActief) repository, which can be inspected at http://media.cwi.nl/sesame/11. The RDF specification is transformed to an associative network structure, called the Knowledge Base. The Knowledge Base contains the knowledge the VMG has on the artworks and their interrelations. Besides the knowledge on the artworks, the VMG maintains a user profile, called the Visitor Model. The Visitor Model incorporates which artworks the visitor has already visited, as well as the interests the visitor had in those. Based on both knowledge sources, the VMG selects a suitable artwork and displays it for the visitor, together with some background information on the artwork.

![Flowchart of the Virtual Museum Guide](image)

Figure 5.1. A flowchart of the Virtual Museum Guide.

KNOWLEDGE BASE

To be able to recommend a sequence of paintings based on their similarity, we chose to represent the artworks in the Knowledge Base by an associative network structure (Collins & Loftus, 1975; Quillian, 1968). The association values in the network indicate similarity: The stronger the association between two artworks, the more similar they are. The associative values in the network are based on the frequency statistics of the keywords that occur in the RDF specifications of the artworks. The general idea is that two paintings of which the keywords greatly overlap might be considered similar to one another.

To determine the similarity of artworks, we applied Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997; Landauer, Foltz, &
Laham, 1998) on the scaled frequency vectors representing the artworks. For this we used the standard TF-IDF weighting scheme (Salton & McGill, 1983), which scales the frequency of terms in a document by the number of documents in which the terms occur. In the VMG’s knowledge base, this means that the frequency of keywords in all RDF specifications is taken into account.

At this point, it is important to note that LSA is more than just a correlation of frequency counts (Deerwester et al., 1990; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). Instead, LSA depends on a mathematical analysis (singular value decomposition) that is capable of a higher-order inference. For example, let us assume that the specification of Rembrandt’s *The Night Watch* contains the key word *claire-obscure*, and Gerard van Honthorst’s *The Merry Fiddler* contains the key word *caravaggists*. The human museum guide knows that these two artworks are similar, because of the two terms are used in similar cultural-historic contexts.12

If *caravaggists* and *claire-obscure* co-occur in similar, yet other, RDF specifications, LSA is capable of making a similar inference as the human museum guide. For example, the RDF specification of Dirck van Baburen’s *Prometheus Being Chained by Vulcan* might mention the keywords *caravaggists*, *light*, *dark*, and *contrast*, and the RDF specification of Rembrandt’s *Ecce Homo* might contain the keywords *light*, *dark*, *contrast*, and *claire-obscure*. Because both *claire-obscure* and *caravaggists* co-occur with the keywords *light*, *dark*, and *contrast*, LSA infers that they are related. In a sense, LSA estimates the likelihood that the word *claire-obscure* would occur in the specification of *The Merry Fiddler*, and the likelihood that *caravaggists* would occur in the specification of *The Night Watch*. For a more detailed, but still non-technical introduction to Latent Semantic Analysis, the reader is referred to Landauer, Foltz, and Laham (1998).

After LSA-values have been computed for all keywords and for all RDF specifications, each RDF specification can be represented by a vector of LSA-values for that specification. The similarity between two RDF specifications is computed by calculating the cosine between the vectors (Salton, Wong, & Yang, 1975). The cosine between two vectors represents their angle, which indicates how much the RDF specifications differ.

**VISITOR MODEL**

To select relevant artworks, the long-term, factual knowledge that is stored in the Knowledge Base, is combined with VMG’s knowledge on visitor interests, stored in the Visitor Model. The rationale of the Visitor Model is that a museum guide perceives how the visitor feels about a particular artwork, but forgets these interests over time, similar to declarative chunks. Therefore, the perceived interest (PI) in a particular artwork *i* may be represented by

\[
PI_i = \ln \left( \frac{f_{bi}}{t_i} \right)
\]

in which *t*_i is the time stamp of the presentation of an artwork, and *f*_bi represents the feedback that the visitor provided when he or she was presented with that artwork. If the artwork appeals to the visitor, *f*_bi = 1; if the visitor is not interested in the artwork, *f*_bi = -1. Because visitors encounter each artwork only once, the summation that is present in Equation 5.1 has been left out.

Figure 5.2 presents an example of the dynamics of the perceived interest in two different artworks. When the visitor sees Artwork A, he or she indicates interest. This is represented by the positive PI-value. However, over time, the VMG forgets the attitude that the visitor had towards the artwork, indicated by the decay of PI. In this example, the visitor disliked Artwork B, which is indicated by the negative PI-value. This PI-value also decays, but increases to indicate that the amount of negativity decreases.
The selection of artworks depends on a weighted scheme of visitor interest and similarity between artworks. Thus, for the selection of a new artwork for presentation, the spreading activation from already visited art is computed. Art that is rated as uninteresting spreads negative activation (because $f_{bi} = -1$); art that is rated as interesting spreads positive activation (because $f_{bi} = 1$). In addition, the spreading activation is scaled according to the similarity between artworks. Thus, artworks that are highly similar spread relatively more activation towards each other. Because of the inclusion of the recency component in Equation 5.2, the influence of recently presented artwork is higher than the influence of artwork presented longer ago. These considerations result in the following equation (Equation 5.3), in which $R_i$, represents the relevance of a certain artwork $i$, $P_{ij}$ represents the perceived interest in already presented artworks ($j$), and $S_{ij}$ represents the similarity between artworks $i$ and $j$.

\[
R_i = \sum_j P_{ij} S_{ij} \quad \text{(equation 5.3)}
\]

This equation represents a match between the Visitor Model, represented by the perceived interests, and each artwork that has not yet been presented. The artwork with the highest relevance will be selected next for presentation. Because $P_{ij}$ can be either a positive value or a negative value (depending on the visitor’s feedback), artworks that were considered uninteresting decrease the relevance of related artworks, while artworks that were considered interesting increase the relevance of related work. Thus, the relevance of an artwork will be high if a visitor expressed interest in related artwork, and did not expressed disinterest in related artwork. Similarly, the relevance will be low (that is, negative), if a visitor only expressed disinterest in related artworks.

After an artwork has been selected for presentation, a web page will be generated that contains a digital reproduction of the selected artwork and some information on the artwork (Figure 5.3). These snippets of information are taken from the Rijksmuseum database, to ensure its correctness and relevance.

**Experiment**

In this section, we test whether the combination of an activation-based visitor model and an associative network-like Knowledge Base is useful for art recommendations. To this
end, we have presented people with artworks that were either selected by the VMG or not, and had them indicate how much they appreciated the selection. The idea is that after a training phase, the VMG should be capable of inferring the participants’ interests from the feedback they had already provided. Therefore, artworks that are selected by the VMG should be better aligned with the participants’ interests than when the artworks are randomly selected.

To study this hypothesis, we developed three conditions. The first two conditions used the cognitive model outlined above (they were termed the Knowledge condition and the VMG condition, respectively); in the third condition the artworks were selected randomly (the Random condition). In the Knowledge condition, the model incorporates the factual knowledge of the museum guide (the Knowledge base), but did not take the visitor feedback into account. In the VMG condition, the model weights perceived interests and the similarity between artworks.

The assumption is that user satisfaction in this study will correlate with the feedback that the participants will give on the presented artwork. If a participant is satisfied with a certain art selection algorithm, he or she will give more positive feedback on the artworks than negative feedback. By analyzing the feedback per condition, we can probe the user satisfaction with a particular selection algorithm. We thus hypothesize that the participants will provide more positive feedback in the VMG condition than in the Knowledge condition, and even less in the Random condition.

In addition, the participants had to indicate their agreement with a set of statements for each condition to determine their attitude towards the different art selection algorithms. A second hypothesis is that participants have a more positive attitude towards the VMG condition than to the Knowledge condition, and again the least positive attitude towards the Random condition.

PARTICIPANTS

Twenty-five undergraduates (19 female) from the University of Groningen participated for course credit. The participants’ age range from 19 to 25 (mean of 21.5). All had normal or corrected-to-normal vision. All were proficient speakers of English.

DESIGN & PROCEDURE

The participants started with a training block of twenty art presentations of historical paintings. Of these, five presentations included landscape paintings, five presented still lifes, five were portraits of historical figures, and five presented genre pieces13. The set of training art presentations was the same for all participants, but the order was randomized. The experimental block consisted of three sequences of art presentations, each consisting of ten items. The sequences differ as to the art selection algorithm used. The order of the selection algorithms was different between participants.

13. Genre pieces are paintings that depict scenes from everyday life. The most famous example of this style is The Milkmaid by Johannes Vermeer.
In the Random condition, ten art presentations were randomly selected from the complete ARIA database. In the Knowledge condition, ten semantically related art presentations were selected. In the VMG condition, the ten art presentations were selected based on the semantic similarity and the feedback that participants gave in the training block (that is, the VMG was used to select the artworks).

Each trial consisted of the presentation of an artwork with the accompanying background information, taken from the ARIA database (Figure 5.3). The display also contained feedback buttons that the participants could press to indicate his or her attitude towards the artwork presentation. In the training block, there were two feedback buttons (labeled Interesting and Not interesting). In the experimental block, there were six feedback buttons, ranging from Extremely interesting to Extremely uninteresting.

Each sequence of ten artwork presentations ended with a small questionnaire consisting of six statements. Three statements related to the associative nature of the sequence of artworks, while the other three statements related to the personalization aspects. The statements were adapted from a questionnaire on usability aspects of an artwork recommender (Cramer, Evers, Ramlal, van Someren, Rutledge, Stash, Aroyo, & Wielinga, 2008). The participants could indicate their agreement with the statements on 6-point Likert-type scales, ranging from Very strongly agree to Very strongly disagree (0-5). The statements are provided in Table 5.1.

The participants were tested in pairs, and were allowed to take as much time as needed to read the information and study the artwork.

<table>
<thead>
<tr>
<th>Type</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Associative</td>
<td>The artworks that this museum guide selects relate to each other.</td>
</tr>
<tr>
<td>2. Associative</td>
<td>I think that this museum guide has no consistent story line in mind when selecting artworks.</td>
</tr>
<tr>
<td>3. Associative</td>
<td>I think the artworks that this museum guide selects form a coherent sequence.</td>
</tr>
<tr>
<td>4. Personalization</td>
<td>I think this museum guide does not understand why I like certain artworks I rated as interesting.</td>
</tr>
<tr>
<td>5. Personalization</td>
<td>I think that the artworks that this museum guide selects correspond to my art interests.</td>
</tr>
<tr>
<td>6. Personalization</td>
<td>The artworks that this museum guide selects do not interest me.</td>
</tr>
</tbody>
</table>

Table 5.1

RESULTS

One participant was excluded for an excessive negative attitude (all responses were in the two most negative feedback options, of which 90% was in the single most negative feedback option). The feedback per condition of the remaining participants is presented in Figure 5.4. An analysis of variance did not show any effect of our condition manipulation ($m_{VMG}=2.49; m_{Knowledge}=2.40; m_{Random}=2.48; F(2,23) < 1$). The same general result was obtained when analyzing the statements. Figure 5.5 presents the scores on the Likert scales for each statement. For this analysis, the scores on Statements 2, 4, and 6 are inverted. Therefore, higher scores reflect a more positive attitude towards the experimental manipulations than lower scores. For analysis, we aggregated the values of the three Associative items into one value, as well as the values of the three Personalization items. An analysis of variance with
Condition as factor showed no significant effect on the Likert scores for the Associative items ($m_{\text{VMG}} = 2.51$; $m_{\text{Knowledge}} = 2.28$; $m_{\text{Random}} = 2.50$; $F(2, 23) = 1.65$; $p = 0.20$), and no significant effect on the Likert scores for the Personalization items ($m_{\text{VMG}} = 2.38$; $m_{\text{Knowledge}} = 2.14$; $m_{\text{Random}} = 2.53$; $F(2, 23) = 2.36$; $p = 0.11$).

### Discussion

Following our assumption that feedback reflects user appreciation, we did not find that the participants appreciated the VMG more than a random selection, or a selection in which only factual knowledge of the museum collection is taken into account. The first reason for the lack of appreciation from the VMG users might have to do with the way the information is conveyed. In the experiment, the virtual museum tour consisted of a sequence of HTML pages containing an image of the artwork and some textual information on the artwork and the artist. One reason why the results of our study are not as expected might have to do with explainability (Cramer et al., 2008). The study by Cramer et al. suggest that users of art recommenders prefer explanations of why a certain recommendation is given. Since this extra layer of explanation is not included in the VMG experiments, this might account for the negative feedback that the participants of the experiment gave.

This issue is illustrated by Figure 5.6. In this example, the visitor has given positive feedback to *The Night Watch*. The visitor has provided negative feedback to the chalk drawing *Resurrection* by Lucas van Leyden. If the visitor is now presented with *Ecce Homo*, a chalk drawing by Rembrandt, this may seem as an incorrect recommendation, because of the similarity between the disliked artwork (*Resurrection*) and *Ecce Homo*. A “reason-giving” extension the VMG that explains that this recommendation follows from the visitor’s perceived interest in Rembrandt’s work may increase the visitor’s appreciation. One option of providing this extension may be by reasoning over the sequence of artworks (cf. Bocconi, 2006), the recommender system might offer meaningful and connective extra information on the overlap...
between consecutive artworks (e.g., Falkovych, Cena, & Nack, 2006; Stock, Zancanaro, Busetta, Callaway, Kruger, Kruppa, Kuflik, Not, & Rocchi, 2007). One recommender system in the cultural heritage domain that seems to have implemented this approach is PEACH (Stock et al., 2007), in which personalization is included by having the user select one of four guides at the start of the tour, each with his or her own focus on the works presented.

**DISCUSSION AND CONCLUSION**

**RELATED WORK**

The key features of the VMG are the combination of the spreading activation network structure of the knowledge base combined with the decaying level of visitor interest. Also, the generation of the knowledge base using Latent Semantic Analysis is an important aspect, as well as the dynamic generation of web content.

Although most of these features have been applied in previous information presentation tools for the museum domain, the combination we apply is, to our knowledge, unique. Also, most other applications focus on the presentation aspects of dynamically generated content, especially in the context of a real, non-virtual, museum, where the mobility of the visitors poses specific challenges for the presentation of information (Hatala & Wakkary, 2005; for a review see Raptis, Tselios, & Avouris, 2005; e.g., Stock et al., 2007). A third obvious difference between related work and our approach is that while most applications focus on the personalized presentation of background information with an artifact, personalization in the VMG involves the selection of the museum artifacts themselves. In this section, we will discuss two systems that seem to be most similar to ours in the key features we have identified for the VMG. That is, both systems – ec(h)o (Hatala & Wakkary, 2005) and PEACH (Stock et al., 2007) – are constructed around a conceptual network, in which selection of concepts is mediated by expressed visitor interests.

Similar to VMG, PEACH (Stock et al., 2007) also adopts an activation-based network. Since PEACH’s main output modality is video, the nodes in the network represent video segments, and the edges represent semantic relations between these video segments. Interest expressed in one video segment propagates as activation through the network to all
related other segments, and new information will be presented based on the activation values of all video segments. This seems to be a similar approach as the VMG deploys, although the level of semantic relatedness in PEACH is less fine-grained, since the relations in PEACH are hand-coded. In the VMG on the other hand, the Latent Semantic Analysis performed on the keywords that represent the artworks ensure that also unexpected – yet relevant – relations may be present.

PEACH also differs from the VMG in the temporal aspects of the relevance feedback. Visitor’s expressed interest in a video segment in PEACH does not extend to another artifact, but only applies to the current artwork. Therefore, decay of visitor interest values is unnecessary. Since the VMG is intended for the dynamic selection of artworks, visitor interest must extend to other artworks.

Just like the VMG, ec(h)o (Hatala & Wakkary, 2005) uses a conceptual ontology as a knowledge base. In ec(h)o, the ontology is based on the Conceptual Reference Model (Crofts, Doerr, & Gill, 2003), which is specifically developed for cultural heritage concepts. Selection of information is subsequently established by reasoning over the relationships in the ontology. ec(h)o also has a decay mechanism to ensure that interests that are more recent are more important than older ones. The mechanism implemented in ec(h)o is however not time-based (as is the decay mechanism of the VMG), but rather the interest values of concepts are normalized such that the highest value stays under a certain upper bound. An advantage of that approach could be that a longer visit to an artwork does not result in ‘forgetting’ of interests, which is a side effect of the way interest decay is modeled in the VMG.

The ec(h)o system differs from the VMG and PEACH in the way relevance feedback can be expressed. Were VMG and PEACH adopt an explicit strategy in which interest as well as disinterest can be expressed, ec(h)o presents the user with three small audio snippets, from which the visitor can choose. The assumption is that the visitor chooses the audio fragment that is the most interesting to him or her. As a result of this design choice, visitors cannot express disinterest. Moreover, they have to base their decision on a small snippet of the actual information, and could well change their minds after they hear all the information. In this sense, ec(h)o does not really incorporate a relevance feedback mechanism.

**CONCLUSIONS**

This chapter describes an recommender system for sequences of artwork presentations, in which the decision on which artwork to present next depends (1) the semantic similarity between the artwork descriptions, and (2) the feedback that users of the system provide at each artwork presentation. The recommender system, the Virtual Museum Guide or VMG, uses principles from cognitive science to provide recommendations that are similar to actual museum guide behavior. In developing the VMG, we ensured that both the visitor’s interests are considered (maintained by the Visitor Model), as well as the relationships between artworks (stored in the Knowledge Base). In the context of a museum, both aspects are important. Because of the educational role of museums, recommending artwork is more than mapping visitor interest on the museum’s collection. The museum needs to ensure that the resulting sequence of artworks is coherent and transfers (part of) the museum’s message. It seems that the Virtual Museum Guide ensures both aspects in artwork recommendation.

An experiment that tested whether users would assess the VMG’s recommendations more positively than randomly selected recommendations did not show an beneficial effect of our current implementation. Two aspects that may relate to this will be further scrutinized in this
thesis. In Chapter 6, we will study if a fine-grained representation of interest may be useful in recommendation of art. While the VMG used explicit feedback buttons, the way feedback may be provided can be very diverse, ranging from the simple button presses used here to more unobtrusive methods, including the time spent observing the artwork (e.g., Claypool, Brown, Le, & Waseda, 2001) or eye gaze analysis. This last option will be explored in Chapter 6.

Another assumption that we implicitly made in developing the VMG, is that the interest that museum visitors expose in art is mainly related to the cultural-historical value of the artwork. Thus, we implicitly assumed that visitors would have a similar interest in artworks that had similar cultural-historical values. This means that we ignored other aspects of the artworks that also may have influenced visitor interest. For example, the use of certain colors, painting or crafting techniques, or a particular arrangement of figures in the scene may also contribute to the appeal a work of art may have to a visitor. These aspects relate to the expressiveness of an artwork (Arnheim, 1954/1974), which is not easily described. For this reason, we turn in Chapter 7 to the domain of scientific literature search, in which the relation between interest and semantic similarity is clearer.
Gaze-based personalization for the presentation of cultural heritage

INTRODUCTION

THE NEED FOR PERSONALIZATION IN CULTURAL HERITAGE INFORMATION PRESENTATION

Consider visiting a cultural heritage museum accompanied by a professional, educated museum guide. The guide has much more information to share about each object in the exhibit than the visitor can, or is willing to, grasp in the time available. Thus, the guide has to make a selection from his extensive knowledge of the museum’s collection that applies to the artwork currently being attended. At the same time, the visitor expects the guide to provide an interesting and appealing tour. Thus, a good museum guide provides an entertaining tour, in which he adapts the information on each exhibited work of art to the perceived interests of the visitors. A system that provides visitors with a personalized tour should ideally incorporate this, what we could call, interest-based entertainment aspect.

On the other hand, cultural heritage institutions also serve an explicit educational role in modern day society. Often, museums are funded by public means, and have as objective to educate visitors on a particular artist, era, or style. Therefore, an ideal museum guide balances the interests of the visitors, who expect an entertaining visit, and the interests of the museum, that wants to educate the visitors (Bell, 2002).

Besides the obvious advantages of personalization from an entertainment point of view, personalization of a learning experience also offers advantages from an educational stance. For instance, providing extra information that matches the interests of the visitor will extend or even deepen the knowledge visitors gain during the museum visit (e.g., Hsi & Fait, 2005; Stock et al., 2007). At the same time, this might also lengthen the time visitors are willing to spend at a certain exhibit. This is particularly important since the average time spent at each object on display is estimated at about only 30s (Beer, 1987; Cone & Kendall, 1978), which is hardly enough to communicate the bare facts of an exhibited artwork, let alone provide interesting extra information. Therefore, also personalization of information presentation within the scope of a single exhibit seems worthwhile.

One way for a museum guide to probe the visitor’s interests is to follow his or her gaze on the work of art. A good museum guide will notice the visitors’ gaze, and will adapt his story accordingly. For instance, a long fixation of one of the visitors to a house in the background of a painting might trigger the museum guide to tell more about the purpose of that house for the painting, or about the architecture from the depicted period.

In this paper, we will exploit this feature of human attention to develop a personalized storyteller for cultural heritage presentations, which will be referred to as Gaze-based Personalization for Art or GPA system. First, we will review work that establishes the relation between eye gaze and interest. Next, we will introduce the GPA system, which will be evaluated in a laboratory experiment to demonstrate that adapting stories told at cultural heritage presentations to the personal interests of visitors improves their experience. We will finish with reviewing related work, and by suggesting extensions and possible further applications of our system.
Eye tracking is a technique that has been available for over 40 years. Many aspects of the eye can be measured, including position, pupil dilation, saccadic movements, and fixation duration (Toet, 2006). Essentially, by tracking (one of) the eyes with a camera that is fixated with respect to the position of the head, eye movements can be singled out from head movements. New computational techniques enable eye tracking without fixating the camera position with respect to the head. Instead, the head position is estimated using pattern analysis of whole-head video sequences, and the eye movements are computed using the head position estimates (e.g., Babcock & Pelz, 2004; Boening, Bartl, Dera, Bardins, Schneider, & Brandt, 2006; Li, Babcock, & Parkhurst, 2006).

Eye gaze is an indicator of overt attention

The direction of gaze and attention correlate to a high degree. Therefore, eye gaze can be regarded as an indicator of attention (Henderson, 2003). Although it has been known for a long time that fixations and attention may deviate (Posner, 1980), during free viewing of natural scenes they are very likely to align (Findlay & Gilchrist, 2005). In this sense, eye movements can be regarded as a behavioral indicator of the spatial allocation of attention.

In addition to the correlation between eye gaze and attention, eye gaze and informativeness are also related (Henderson & Hollingworth, 1999). Several studies (Antes, 1974; Loftus & Mackworth, 1978; Mackworth & Morandi, 1967; Yarbus, 1967) established the relationship between fixations on different regions of a picture and the informativeness of these regions. An interesting aspect of this relationship is that the fixation patterns are directly related to the goal that the participants have. For example, a request to memorize a scene results in different fixation patterns than viewing the painting with the goal of determining the wealth of the people depicted (Yarbus, 1967). Mackworth and Morandi (1967) and Antes (1974) based their measure of the informativeness of a region on ratings that users gave to the different regions. Thus, they established a relationship between eye gaze and users’ idea of informativeness. Under the assumption that when perceiving an artwork in a museum, a visitor’s goal is to find those regions on a painting that are most interesting to him or her, these studies suggest that eye gaze is an indicator of interest that museum visitors might have in specific regions of realistic paintings.

Various researchers (e.g., Chen & Zelinsky, 2006; R. M. Cooper, 1974; Huettig & Altmann, 2005; Yee & Sedivy, 2006) have shown the relationship between semantic content and eye gaze. In an experiment that involved eye fixations and spoken words, Cooper (1974) showed that people fixate more on regions of interest (ROIs) on the screen that they hear in a snippet of spoken text than on ROIs that are not mentioned in the spoken text. In addition, ROIs that bear a semantic relation to words mentioned in the spoken text are fixated more than ROIs that are unrelated to the text. These results demonstrate that eye gaze is also mediated by the semantic content of the display. Again, if we assume that one of the goals of museum visitors is to find interesting aspects on each work of art on display, then a prolonged gaze on a certain object in the painting may be the result of the overlap between the semantic content and the interest of the visitor.

Eye gaze as a pointing device

In many Human-Computer Interaction applications, gaze has been deployed as an explicit pointing device, much like a computer mouse. Typically, users may interact with an...
interface using stares or blinks to indicate that an action needs to be performed (e.g., Hornof & Cavender, 2005; Jacob, 1991). Although the user interaction is mediated by gaze in our application as well, it explicitly differs from these applications because we do not intend users to control the presentation of information with their gaze, but rather analyze the free viewing behavior to determine the most interesting regions of an artwork.

The above reviewed studies suggest that it is possible to develop a system that personalizes background information on the basis of eye gaze when presenting a work of art. In what follows, we will discuss how we implemented this idea, and we will present an experiment in which that implementation is put to the test.

GAZE-BASED PERSONALIZATION FOR ART

The Gaze-based Personalization for Art (GPA) system that we propose uses the eyes’ gaze as an indicator of interest, and presents information on paintings based on the perceived interests of the user. The information about the artworks is presented using natural speech. Because the selection of information depends on the direction of gaze, which differs between individuals, the story that is told at each artwork also differs between individuals.

The sum of the duration of the various fixations on different regions of interest (ROIs) on a painting determines the choices made by the GPA system. This summation will be referred to as gaze duration.

ANNOTATING ARTWORKS

Before interesting audio snippets can be selected, each painting has to be annotated. That is, the spatial dimensions of every ROI on a painting that might be of interest to a virtual museum visitor of the system have to be identified and related to an interesting piece of information. In GPA, the spatial dimensions of the ROIs are represented in a content map that indicates which pixels on the screen belong to which ROI (Figure 6.1). Regions of interest (ROIs) on the content map can be anything: a figure or a group of figures, objects, buildings, animals, et cetera. In addition, ROIs may consist of multiple components: Several smaller areas on the display can combine to one ROI. An example of this is if there are multiple animals on the artwork, and the audio snippet explains something about the use of animals in paintings of a certain genre.

Figure 6.1. Content map of the painting “The Sacrifice of Iphigenia” by Jan Steen. In this content map, ten regions of interest (ROIs) are identified on which information can be presented.
Because ROIs may differ in size, bigger ROIs would have a higher probability of being attended than smaller ones, if there would be no control of eye movements. However, the control of eye movements is accurate enough to ensure that most of the fixations are on the intended ROIs. We therefore assumed that the number of fixations that land on an unintended ROI would not significantly influence ROI selection.

**DETERMINATION OF INTEREST**

The choice of which audio snippet to present depends on the recorded fixations. The recorded fixations are represented by a fixation map (Velichkovsky, Pomplun, & Rieser, 1996; Wooding, 2002), which is a two-dimensional array in which the gaze duration on a specific region of the screen is represented. Depending on the need for spatial resolution of the fixations, the dimensions of the array could vary between the pixel dimensions of the display – providing maximum spatial resolution – or one or more orders of magnitude smaller than that. Because in the context of the GPA system we want to identify which ROIs on a painting are being fixated, we are only interested in whether fixations are within the boundaries of these ROIs. Since in our study the average ROI size is relatively large (43 pixels in diameter, see also Figure 6.1), we set the size of a patch on the fixation map at 10 x 10 pixels. The pixel dimensions of the display we used in the interest-aware system were 1024 x 768, and therefore the fixation map consisted of 102 x 77 patches of 10 x 10 pixels (with the patches on the right side of the display 8 pixels wide, and the patches on the bottom of the display 14 pixels high). Because the number of fixations is typically very large, we followed Wooding’s (2002) suggestion that each fixation could be represented by the fixation location only, instead of a Gaussian distribution reflecting noise in the eye gaze recording. It is unlikely that this simplification will result in different information selection, because the ROIs we defined were much larger than the variance in the Gaussians.

Over time, the durations of new fixations are added to the already accumulated values of the fixation map. In the interest-aware system, the updates to the fixation map are made after successive presentations of audio snippets. Thus, the selection of information is based on a user’s gaze from the initial presentation of a painting up until the moment of information selection.

**REGION OF INTEREST SELECTION**

The interest that people show in the ROIs on the painting is estimated with the following equation:

\[
I_i = \sum D \cdot C_i \text{ with } C_i = \begin{cases} 
1 & \text{if } c = i \text{ with } c \in C \\
0 & \text{otherwise}
\end{cases}
\]  

\text{ (equation 6.1)}

\(I_i\) indicates the interest in ROI \(i\), \(D\) indicates the array representing the fixation map, and \(C_i\) indicates the array representing which cells belong to ROI \(i\). \(C_i\) can be derived from the original content map by substituting all cell values (\(c\) in Equation 6.1) by 1 if and only if they belong to ROI \(i\), and by zero otherwise.

By combining the fixation map and the content map, we arrive at a set of gaze durations representing individual interest for each ROI on the screen. Based on these values the selection of information can be performed by selecting the ROI with the highest interest that has not been selected before, and presenting the associated audio snippet.

In developing the GPA system, we experimented with different algorithms that estimated the probability that a fixation on a certain ROI was the result of intentionally looking at that.
ROI, instead of by other factors, such as saliency differences between various regions on the painting (Itti & Koch, 2001; Itti, Koch, & Niebur, 1998; Koch & Ullman, 1985), or saccadic noise (Kowler & Blaser, 1995). Many studies show that, besides attention-mediated processes, eye gaze is also controlled by the physical properties of whatever a person is watching (e.g., Henderson, 2003; Kootstra, Nederveen, & De Boer, 2008; Theeuwes, 1992). The typical distinction is between top-down control of eye-movements, in which some cognitive operations are performed, and bottom-up control, which refers to eye movements caused by features of the visual field. Bottom-up control is usually linked to saliency, a concept representing the differences in physical properties of the visual field, such as color, contrast, orientation, or motion. A study by Theeuwes (1992) demonstrates that if a highly salient distractor is presented in a visual search task, participants make an eye movement to this distractor before fixating on the search target. Therefore, given that the visual acuity of a region on a painting is high enough, it will cause the eyes to fixate there. However, when freely viewing natural scenes (such as classical paintings), the influence of saliency on the fixation duration is typically much less than the influence of top-down modulating factors (such as informativeness of the region, or interestingness, Henderson & Hollingworth, 1999). Thus, it is interesting to see whether in the context of artworks, the saliency of an artwork contributes to the direction of gaze, and whether saliency should be considered when developing gaze-based personalization systems.

PRESENTING INFORMATION USING SPEECH

When the most interesting ROI is selected, the associated audio snippet can be presented. Obviously, when a subsequent ROI needs to be selected, all already presented ROIs are excluded from selection. Care was taken in the construction of the audio snippets that they only referred to a single ROI, thus diminishing the chance that gazes to other ROIs could be caused by reference to the semantic content of those ROIs (R. M. Cooper, 1974; Huettig & Altmann, 2005; Yee & Sedivy, 2006).

We chose to present the users of the interest-aware system with spoken text. Using another modality than vision for information presentation is useful in the context of attention-aware systems, because the eyes are not distracted by extra visual information that otherwise might appear on the screen (cf., Starker & Bolt, 1990).

VALIDATION OF GAZE-BASED PERSONALIZATION FOR ART

We performed a laboratory experiment to study whether the GPA system outlined in the previous section would prolong user interest when perceiving art. In this study, participants were asked to look freely at a series of paintings while a voice-over gave information on these paintings, and to indicate for each painting when they lost interest in that particular painting. We hypothesize that when participants receive information on ROIs they implicitly expressed interested in with their gaze, they would stay interested longer and indicate their loss of interest later.

EXPERIMENTAL MANIPULATIONS

To contrast the behavior of participants interacting with the actual GPA system, we designed three information-selection conditions. In the first condition, henceforth referred to as the maximum condition, the participants received information via audio snippets for ROIs in which they showed maximum interest. The maximum condition is contrasted with two control
conditions: The random condition and the minimum condition. In the random condition, participants were presented with the information in a random order, not depending on their gaze durations; in the minimum condition, participants received information pertaining to the ROIs for which they expressed the least interest (that is, the ROIs with the lowest gaze durations). Our main hypothesis is that viewing time on a painting is mediated by these information-selection conditions: Longest for the maximum condition, in between for the random condition, and shortest for the minimum condition. Note that the random condition can be seen as a standard online guide in that the order in which information was presented to the participants was not modulated by knowledge of the participant’s perceived interests.

A secondary hypothesis will be that in the maximum condition users will look longer to the ROIs on which they receive information than in the random and minimum conditions. This effect is hypothesized to result from the mapping between the users’ interest and the presented information. The rationale of this hypothesis is as follows: Given the assumptions that eye gaze follows interest, and that eye gaze is influenced by the semantic content of presented audio, the gaze durations on selected ROIs in the maximum condition would benefit from both factors, whereas the gaze durations on selected ROIs in the random and minimum condition would only benefit from the latter factor. Due to a random presentation order of the information in the random condition, we expect the gaze duration on presented ROI in the random condition to be in between that in the maximum and minimum conditions. This reflects the probability that a participant is presented with an audio snippet that relates to a highly attended ROI.

PARTICIPANTS

Thirty undergraduate students from the University of Groningen participated in the study. The mean age of the participants was 22.6 years. All participants had normal or correct-to-normal vision and normal hearing and received course credits for their participation. All were native speakers of Dutch.

APPARATUS

The paintings were displayed on a 19” CRT monitor. We measured eye gaze with the Eyelink I eye tracker from SR Research. The Eyelink I is a head-mounted eye tracker with a spatial resolution of 0.01°, a sampling rate of 250Hz, and an average gaze position error of 0.5°-1.0°. Note that the average ROI size is 1.2° (an average diameter of 1.7 cm at approximately 80 cm distance). The Eyelink I delivers pupil position and head marker position in real-time (based on four LED-arrays attached to the screen), and can thus account for both eye movements and (small) head movements. Although the Eyelink I is a head-mounted system, the strain on the participants was acceptable. We therefore do not expect differences in viewing behavior as compared to natural viewing behavior.

STIMULI

Paintings

To stay as close as possible to a museum setting, whilst still in a controlled laboratory environment, we used digital reproductions of paintings from the Rijksmuseum collection, with a resolution of 1024 x 786 pixels. The selection of paintings was such that they contained enough identifiable elements to present sufficient audio snippets. The paintings either depicted a biblical or mythological scene, or depicted a scene from everyday live. (Genre
painting). Most selected artworks were created during the Dutch Golden Age (roughly the 17th century). Figure 6.2 presents examples of the paintings used in the experiment.

![Figure 6.2. Four examples of paintings used in the experiment. (a) “The Sacrifice of Iphigeneia” by Jan Steen (1671) (b) “Prince’s Day” by Jan Steen (c. 1665) (c) “The Fall of Man” by Cornelis van Haarlem (1592) (d) “Aeneas at the Court of Latinus” by Ferdinand Bol (c. 1661-63).]

Audio snippets

Each painting was accompanied by a sequence of audio snippets. The first, introductory audio snippet had an average duration of 20s (±1.5s). It contained the name of the artist, the year the painting was produced, a comment on technique or a small anecdote regarding the painting as a whole or the artist, and finally the title of the painting. The introductory audio snippets were the same for every participant. We made sure that no direct references to ROIs on the painting were made, except when this occurred because of the painting’s title. After the introductory audio snippet a maximum of 10 other audio snippets were presented during 10s each (±1.5s), separated by one-second intervals. The short intervals between the audio snippets ensured that a sequence of audio snippets sounded naturally. The texts of the audio snippets were based on the explanations that accompanied the paintings as found on the website of the Rijksmuseum (www.rijksmuseum.nl). We adapted the texts so they would be self-contained and would fit in a timeframe of 10s. As an example, Appendix A presents the text of the audio snippets for the painting “The Sacrifice of Iphigeneia”.

DESIGN AND PROCEDURE

Every participant was tested individually. First, the eye tracker was set up and calibrated, and the participants were instructed on the remainder of the experiment. Second, participants were presented with 26 paintings while hearing a sequence of audio snippets accompanied each painting. Each presentation of a painting constituted one trial. Each trial was preceded by a drift correction screen. The first painting was a practice trial, and was always the same for all participants, as was the sequence of audio snippets presented during this trial.

The participants were instructed that they could indicate loss of interest in the painting or the information by pressing the return key, which would result in a new painting being presented. Pressing return signaled the participants’ response.

After the practice trial, 25 paintings were presented in pseudo-random order to each participant. Initially, we also were interested in the effects of saliency on the selection of audio-snippets. However, since there were no effects from the saliency manipulation, we collapsed those conditions with the default conditions. Therefore, ten paintings were presented in the
maximum condition, ten were presented in the minimum condition, and five were presented in the random condition. To counterbalance possible effects of fatigue or loss of interest across participants, each painting was presented as often in the first half of the presentation sequences as in the second half of the presentation sequences. Moreover, every painting was presented equally often in the maximum as in the minimum condition.

If a participant pressed the RETURN key, the voice-over finished the current audio snippet, and then moved on to the next trial. If participants listened to all audio snippets of a trial, the system also progressed to the next trial. The set of audio snippets for a painting was the same in every condition; only the order in which they were selected differed, depending on the condition and the gaze of the participant.

RESULTS

As the main dependent variable is the listening duration, two participants were excluded from further analyses as they listened to only a single audio snippet in more than 50% of the trials, and three participants were removed because they listened to all audio snippets in more than half of the cases. This leaves 25 participants for further analysis, of which the average listening duration in the initial trial was 69s (SD=42s) and in the experimental trials 76s (SD=31s).

Although the main analysis of interest is the effects of different audio-selection conditions on listening duration, we will first turn to the effects of saliency. The saliency condition was included as the assumption was that saliency could influence eye movements independent of actual interest. However, no differences were found related to the saliency condition (F(1,24)<1) or to the interaction of saliency with audio-selection condition (F(1,24)<1). We have therefore collapsed the data over Saliency for all subsequent analyses.

In contrast to our hypothesis, there was no effect of Audio-selection on listening duration (means: m maximum=76.5s, m minimum=74.3s, m random=76.8s, F(2,24) = 1.01; p=0.37). Additional exploratory analyses including different personal characteristics (e.g., age, gender, reported interest in art, etc) or performance measures (e.g., ratio of gazes on ROIs) as co-variates did not change this outcome.

Based on the assumption that interest is reflected by increased gaze durations on the regions of interest, the second hypothesis was that participants look longer to items on the artwork they are interested in than to items they are not interested in. To test this hypothesis, we studied the gaze duration on ROIs during the presentation of audio snippets.

Gaze durations before the presentation of an audio snippet were used to select audio snippets in the maximum (and minimum) condition. Thus, the second hypothesis is that during the presentation of audio snippets, gaze durations on the ROIs associated with the presented audio-snippet should be longer in the maximum condition than in the random or minimum conditions. Furthermore, to rule out the possibility that the participants had not seen the ROI under consideration, we excluded all instances in which the critical ROI was not fixated before the presentation of an audio snippet (this led to one participant being excluded that had not fixated the critical ROIs in one condition).

Figure 6.3a presents average gaze durations on the presented ROIs during the presentation of the associated audio snippets. The gaze durations between conditions differ significantly (F(2,23) = 1.10; p<0.001). This may be a first indication that participants will remain interest in ROIs that they gaze at irrespective of the presented audio information. If gaze was primarily determined by the content of the audio snippet (R. M. Cooper, 1974), we would have expected no difference between the bars in Figure 6.3a.
A second indication that gaze duration reflects interest is presented in Figure 6.3b. If prior interest guides the eye movements, then the ROIs that the participants previously gazed at should be gazed at again, irrespective of the presentation of an audio snippet. However, if the audio snippet determines the direction of gaze, then we would expect a difference between the gaze durations on the most attended ROIs during presentation of the audio snippet. This is because for the random and minimum condition, the presented audio snippet is not about the most attended ROIs. Thus, gaze would be expected to deviate from these regions. However, an analysis of variance did not show a difference in gaze duration in Bars IV, V, and VI (means: max = 2064ms, min = 2219ms, rand = 2167ms, F(2,23) <1). We did not find evidence that the content of the audio snippet determines which ROIs will be gazed at more, suggesting that interest is a stronger determinant of eye gaze than the content of an audio snippet.

**DISCUSSION**

In this experiment, we demonstrated the role of eye gaze in interest detection. Regions on a painting that are selected because they are being attended are gazed upon to a larger extent than regions that are selected because they were not or hardly attended (Figure 6.3a). Interestingly, this observation does not result in longer listening durations (the total time spent viewing each artwork). This may be in line with earlier observations that museum visitors have a typical dwell-time they are willing to spend at each exhibited art work (Beer, 1987; Cone & Kendall, 1978). Thus, assuming that the time people are willing to invest in each art work on display is roughly constant, as well as limited, the need for a smart ordering of the information that the museum wants to convey becomes prevalent.

One complicating factor in the interpretation of the results of our study is that eye gaze is implicated in multiple behavioral mechanisms. Thus, other factors could have influenced eye gaze. In the “Regions of Interest Selection” Section, we already mentioned the possible influence of saliency. Areas on the painting that stand out on a certain dimension such as color or contrast are more likely to be attended. However, when viewing paintings, the influence of saliency is much less than the influence of interest (Henderson & Hollingworth, 1999). Moreover, we counterbalanced the artworks in such a way that each artwork occurred in every condition equally often.

In addition, it has been shown that gaze is also affected by the semantic content of an aurally presented message (R. M. Cooper, 1974). This could have affected the results of our study as well. However, Cooper found that fixation durations were increased on ROIs that were mentioned in the message. If the eye gaze of the participants in our study was mainly driven by the content of the audio snippets, we would therefore find a difference in gazes on the most
attended ROIs in the maximum and the minimum condition. The fact that we do not find this
difference (e.g., Figure 6.3b) can be seen as evidence that in our study, the gaze durations are
not driven by the semantic content of the audio snippet. One interpretation of this difference
between Cooper’s study and ours is that in his study, the ROIs were well-defined and known
to the participants in advance (because of the clear arrangement of items on the display),
whereas in our study, the demarcations of ROIs are less clear, due to the natural scenes
depicted. Moreover, the participants were unaware of the existence of ROIs, and may thus be
less likely to acknowledge auditory references to the depicted items by a quick fixation.

**DISCUSSION AND CONCLUSION**

In this paper, we proposed a personalized system for information presentation at art
exhibitions, termed Gaze-Based Personalization for Art. The system uses point of gaze to infer
a visitor’s interest, following prior studies that suggest a correlation between eye gaze and
attention. We determined regions of interest (ROIs) for the artworks, which together form
content maps of the artworks. By combining the content maps with dynamically updated
fixation maps, we can compute the how much each ROI is fixated, and select the most fixated
ROIs for information presentation.

In an experiment, we contrasted the GPA system with two control systems that had a
random selection mechanism and a negative selection mechanism, respectively. The results
show that presented information on the most attended ROIs increases the fixation duration
on those ROIs, although the total time spend examining an artwork does not seem to be
influenced by personalized information presentation.

Our results suggest that if visual perception of the artwork is important when presenting
information, gaze-based ordering of the to-be-presented information is useful. For instance,
when the information relates to a specific use of colors (e.g., the use of dark and light colors
in the paintings by Rembrandt van Rijn), it is important that the museum visitor attends the
regions of the painting on which that technique is exposed. The experiment discussed above
suggests that if the visitors are not attending those regions in the first place, they are not likely
to attend them when the presented information is about the use of light and dark in Rembrandt’s
paintings. Therefore, our work may help cultural heritage institutions to adapt the order of the
information that they want to convey to optimize the knowledge transfer to the visitor.

**RELATED WORK**

**Personalization of cultural heritage presentation**

The idea of personalizing certain aspects of a cultural heritage experience has been
studied extensively. Most applications focus on the presentation aspects (e.g., Falkovych,
Cena, & Nack, 2006; Hatala & Wakkary, 2005; for a review see Raptis, Tselios, & Avouris,
2005; Sparacino, 2002; Stock et al., 2007). Some work is directed at personalized sequences of
the objects at display (e.g., Fink & Kobsa, 2002; Van Maanen, 2007). All of these approaches
provide personalization of museum content by adapting which information is being presented
to the individual user. To our knowledge, the current study represents the only attempt at
using eye-gaze as an informative devise for personalization of cultural heritage content.

**Gaze-based information selection**

Another relevant concept that has been studied before is the selection of information
using eye gaze. Previous work in the gaze-based selection of information has focused on
explicit information retrieval systems (e.g., Oyekoya & Stentiford, 2007; Puolamäki, Salojärvi,
Savia, Simola, & Kaski, 2005). In the GPA system, the selection of information is implicit, because museum visitors are not intentionally looking at certain regions of the artwork in order to receive information about these regions.

There exists however two systems that share many features with the GPA system. These are iTourist (Qvarfordt & Zhai, 2005) and the gaze-responsive system by Starker and Bolt (1990). Both systems track the user’s point of gaze, and present new information that is selected on the basis of where the user was looking. In iTourist, the information is related to tourist information on a virtual city map. iTourist presented spoken as well as visual information on points of interest on the map. The system by Starker and Bolt presents spoken information on objects on a small planet, similar to iTourist and GPA. These systems differ from GPA in that users of both applications are aware of the manipulatory role of their gaze, due to the particular layout of the display (a city map with marked tourist highlights and a rotating planet with isolated objects). By contrast, users of the GPA system are not aware of the manipulatory role of their gaze. Participants in our experiment were not informed on the reason for measuring their eye movements. In addition, the regions of interest on an artwork are less well-defined than the regions of interest on a city map or on a constructed 3D world, which makes it harder to intentionally fixate a certain region. Therefore, it seems to make more sense to use eye gaze in the cultural heritage domain in a diagnostic fashion rather than an manipulatory or intentional fashion (Duchowski, 2002).

MORE NATURAL SETTINGS

Although we tested the GPA system in a controlled laboratory environment, the approach seems suitable for environments that are more natural as well. For instance, the online presentation of sequences of artwork may be augmented with gaze-based interest awareness (Van Maanen, Janssen, & Van Rijn, 2006). With the advent of online information presentation, cultural heritage institutions are starting to make their collections available online. Many museums have websites displaying digital reproductions of part of their collection. Moreover, recent advances in eye tracking technology as well as increased quality of standard webcams have brought online non-intrusive gaze tracking to the desktop (e.g., the COGAIN initiative, www.cogain.org, or Hansen, Hansen, & Johansen, 2001). Using webcams for eye tracking may enable the use of gaze-based interest awareness to adapt the presentation of information on the artworks available at museum websites.

In addition, more advanced eye tracking devices have been developed that allow for more free movement of the user (e.g., Babcock & Pelz, 2004; Boening et al., 2006; Li, Babcock, & Parkhurst, 2006). With these devices, it becomes possible to freely wander through a museum, while your point of gaze is being tracked. This allows for adaptive information presentation in the real museum, for instance using headphones.

CONCLUSION

Eye gaze may provide useful insights in people’s interest, which can be used in cultural heritage applications. Detecting personal interests of museum visitors enables personalized presentation of the exhibit’s information. This may increase the enjoyment that visitors have when attending an exhibition, but may also improve their learning experience, because the ordering of the presented information may be such that the information aligns with the visitor’s prior knowledge and interest. The GPA thus balances a museum’s educational role and a visitor’s personal interests, just like a good real-life museum guide.
APPENDIX

Example of audio snippet s of “The Sacrifice of Iphigeneia”. 0 indicates introductory snippet, which was presented first. 1-10 indicates the remaining snippets, the order of which may be determined by the participants gaze and the condition.

0. Het volgende schilderij is gemaakt door Jan Steen en dateert uit de periode rond 1671. Jan Steen werd vooral bekend om zijn “genrestukken” met vrolijke gezelschappen. Maar ook met portretten en schilderijen over mythische verhalen. Dit werk is gemaakt met olieverf op doek. De titel van het schilderij is “Het offer van Iphigeneia”.

1. Boven in de rook zit Artemis, de Godin van de jacht. Ze is herkenbaar aan de maansikkel op haar hoofd en de pijl en boog. Artemis werd ook vereenzelvigd met de maangodin Selene.

2. Jan Steen hield zich niet aan de regels van de historische schilderkunst. Om het verhaal voor zijn tijdgenoten herkenbaar te maken, gaf hij de meeste personen geen oud Griekse, maar 17de eeuwse kleding.

3. Een jongetje loopt bedroefd weg op het schilderij. Het is Amor, de god van de vleeselijke liefde. Hij is herkenbaar aan zijn pijl en boog waarmee hij mensen verliefd kon maken.


5. De schilder beeldt het moment af vlak voor dat een offer wordt gebracht. De beul staat al klaar om zijn slachtoffer te doden, het mes glistert in zijn hand en hij kijkt bloeddorstig naar het offer.

6. De geknielde vrouw is met veel zorg afgebeeld, met details als de glanzende kleding en de voeten die vies zijn van het lopen. Het licht van het vuur rond haar afgewende hoofd geeft haar zelfs iets mysterieus.

7. In het midden ziet u Iphigeneia. Omdat het oorlog is zal zij worden geofferd aan de Goden om hen gunstig te stemmen. Het offeren van mensen, dieren en voorwerpen was zeer gebruikelijk in de Griekse oudheid.

8. Rechts bovenin is een vrouw aan het bidden. De gevouwen handen zijn een christelijk gebaar wat niet hoort bij de Griekse oudheid. Door dit wel te gebruiken is het verhaal herkenbaarder voor de tijdgenoten van de schilder.

9. De figuren rechts hebben een exotisch tintje vanwege hun verschillende hoofddeksels. De tulband, de lauwerkrans, een soort bisschopsmijter en een variatie op een Romeinse helm zijn te zien.

10. De afgebeelde koning is Agamemnon uit Griekenland. Hij wil Troje belegeren, maar telkens als hij wil uitvaren is het windstil. Om de boze godin Artemis gunstig te stemmen moet hij zijn dochter offeren.

English translation

The following painting is from Jan Steen and dates from around 1671. Jan Steen became especially well known for his genre paintings with fun-loving, cheerful groups of people. But also with portrait paintings as well as paintings of mythological scenes. This piece is oil on canvas. The title of this painting is “The Sacrifice of Iphigeneia”.

In the upper part of the painting sits Artemis, the Goddess of hunt. She can be
recognized by the crescent moon and her bow and arrow. Artemis was considered the same Goddess as the moon Goddess Selene.

2. Jan Steen did not keep strictly to the rules of history painting. To make the story recognizable for his contemporaries, he has not dressed all the people in ‘Iphigenia’ in classical Greek costumes, but seventeenth-century clothes.

3. A little boy leaves the scene crying. It is Amor, the God of erotic love. He can be recognized by his bow and arrow that he used to make people fall in love.

4. No one dies in this sacrifice. In a miracle, the human victim is switched with a deer. Consequently, only the deer dies.

5. The painter has pictured the moment just before Iphigenia is to be sacrificed. The executioner is about to kill her, the knife shines in his hand while he is looking cruelly at his victim.

6. The kneeling woman has been depicted with great care, with realistic details such as the shine of her silk clothes and her bare feet that have become a little dirty from walking. The light from the wood fire that shines around her turned head gives this woman a mysterious air.

7. In the center stands Iphigeneia. Because of a war, she will be sacrificed to pacify the Gods. Human and animal sacrifices as well as sacrifices of goods were very common in ancient Greece.

8. In the upper right corner, a woman is praying. Her hands are folded in a Christian manner, which is anachronistic for a scene of ancient Greece. By using this, the story has become recognizable for the painter’s contemporaries.

9. The people on the right all have an exotic appearance due to their different headpieces. You can see a turban, a laurel wreath, a miter, and something similar to a roman-style helmet.

10. The depicted king is Agamemnon of Greece. He is determined to lay siege to the city of Troy. Yet, every time he tried to embark his fleet, a calm descended. To sooth the angry Goddess Artemis, he must sacrifice his daughter.
Abstract Recommendations by a Cognitive Model


INTRODUCTION

In cognitive science, there has been a long tradition to perceive human behavior as a form of information processing. Within this tradition, human cognitive processes are seen as operating on similar principles or algorithms as computer programs, since both cognition and computer programs have or have been developed to process information. This view has lead to the birth of Artificial Intelligence as an independent research field (McCarthy, Minsky, Rochester, & Lebiere, 1955), but has also guided the development of cognitive theories (e.g., Anderson & Milson, 1998; Marr, 1982; Newell, 1990). Even today, the apparent functional overlap between artificial computational systems and the human information-processing system is still influential in cognitive theorizing (e.g., Griffiths, Steyvers, & Firl, 2007).

Many cognitive theorists believe that human beings optimize their behavior to successfully cope with the environment (e.g., Anderson, 1990; Marr, 1982; Oaksford & Chater, 1998). This means that, through evolution and learning, human behavior has adapted to be the most suitable behavior in any given circumstance or environment. This is a capacity also desirable in artificial systems design, especially when these systems have to operate on an unknown or dynamic environment. Therefore, computer scientists and artificial intelligence researchers have studied how computer systems can optimize their behavior as well (e.g., Goldberg & Holland, 1988; Kohonen, 2001).

A domain that has not benefited that much from this cross-fertilization is the problem of selecting relevant information, either for oneself or for others. The research field that studies how to disclose relevant information is known as Information Retrieval (Salton & McGill, 1983). A typical field in which the problem of selecting relevant information arises is the scientific community. For example, the number of scientific publications in the relatively small ISI subject category Information Science & Library Science was 2054 in 2006. This means that researchers working in this area have to read (or at least scan through) over two thousand papers a year to keep up with the current developments. However, this number is, if anything, an underestimation of the total number of potentially relevant papers, as this number only holds if the researcher is interested in a single subject area. In practice, most researchers work on the intersection of multiple domains, increasing the number of potentially relevant papers enormously. In general, because of the continuous increase of storage capacity for digital media, and the increased availability of digital or digitized media sources, companies, institutions, and individual people are being confronted with an increase in the amount of information that potentially is relevant to their purposes.

In this paper, we will describe a system that partly solves this problem for the scientific domain: Our system selects relevant scientific papers from a large collection of scientific abstracts. Instead of working from a pure computer science perspective, we will present a system that is based on constraints from cognitive theories. In particular, we chose to follow the rational analysis approach (Anderson, 1990; Oaksford & Chater, 1998), as incorporated in the ACT-R architecture of cognition (Anderson, 2007a). The rational analysis approach states
that human memory is optimally adapted to fit the needs of the environment we live in, based on the interactions of the cognitive agent with the environment in the past. This approach has been successfully applied to predict various aspects of human behavior (e.g., as reviewed by Chater & Oaksford, 1999).

We will begin with an analysis of how users behave when engaging in the selection of information. Next, we will discuss how the ACT-R cognitive architecture incorporates rational analysis, and how this can be applied to information selection. We will continue with an outline of an application based on the resulting model, the Personal Publication Assistant (or Publication PA for short), and how this application behaves under different conditions, as well as a user study that will demonstrate the applicability of our approach in a real world setting. In the last section, we will discuss in what way the Publication PA deviates from other approaches towards the task of matching papers to researchers, or vice versa.

**INFORMATION SELECTION**

An example of the problem addressed in this paper is the selection of relevant information when attending a large, multi-track scientific conference. Often, an attendee finds him or herself overwhelmed by the amount of presentations that can be attended. With so little time to find the talks that are really interesting, changes are that one ends up in the wrong track, listening to presentations that hardly kindle ones interest, while in another track relevant work is being discussed. Although this might bring unforeseen beauty, often a better selection of relevant work would be preferable. There are solutions to this problem. For example, giving the attendees the proceedings well in advance so they have more preparation time. However, this solution is often not viable due to practical constraints. A better solution might be to provide an automatic recommendation based on the personal interests of the conference attendees, which is the approach that will be discussed in this paper.

To build a successful recommendation system, it is important to know how the selection process takes place in unsupported settings. The information selection process starts when a researcher registers at a conference and receives a copy of the conference proceedings. Based on informal analyses, the next step is to perform a quick scan of all titles, author names, or abstracts for words or names that are familiar. If an entry contains enough interesting words, it is selected for further and more careful reading. Obviously, the assumption that is made implicitly, is that individual words in the abstract accurately reflects the contents of the paper or presentation. Ries et al. (Ries, Su, Peterson, Sievert, Patrick, Moxley, & Ries, 2001) have shown that this assumption holds for abstracts and papers, at least in the medical domain. In order to determine if a word qualifies as interesting in the context of the conference, the researcher might assess whether she has used the word in her own research in the past. One could say that the researcher tries to discover the degree of familiarity she has with an abstract, and if that degree of familiarity is high enough, she selects that presentation as potentially worthwhile to visit.

To assist a researcher in the information selection task, we propose a model of the recognition aspects of the task. That is, we propose a model that makes a preselection from the available information based on a notion of familiarity adapted to the individual researcher. To achieve this, we will develop models of the declarative memory systems of individual researchers (henceforth refered to as user models) and of the process of recognizing words. Each user model can be seen as a representation of an individual researcher’s interests, as it incorporates the frequency, recency, and context of the words used by the researcher to
describe her research. In previous research (Anderson & Milson, 1989; Anderson & Schooler, 1991), a formal model has been developed of how the retrieval of declarative facts from memory can be described. In the next section, we will give a detailed overview of that model, but we will highlight the two most important aspects here. One key idea is that declarative memory is optimally adapted to serve the needs of the cognitive agent (Anderson, 1990; Oaksford & Chater, 1998). The other is that most facts in declarative memory are initially formed by perception (Anderson & Schooler, 1991). Combined, this means that the adaptive nature of declarative memory is essentially a reflection of the perceptions of the cognitive agent. As a consequence, this means that looking for structure in the environment can derive the structure of declarative memory.

**RATIONAL ANALYSIS OF MEMORY**

Anderson and Schooler (1991) showed that the probability that a memory will be needed in the near future depends on the pattern of prior exposures to the piece of information stored by that memory. For example, the probability that someone will contact you by email today depends on the frequency and recency of her emails to you in the past (Anderson & Schooler, 1991). Likewise, the probability that you will need some declarative fact from memory right now depends on the frequency and recency of the prior usage of that fact. Both relations are captured by Equation 7.1, in which $B$ stands for the base-level activation (reflecting the probability), $t_i$ stands for the time since exposure to event $i$, and $d$ represents the speed with which the influence of each exposure decays. The summation is over all $n$ previous encounters of the events ($i$).

$$B = \ln \left( \sum_{i=1}^{n} t_i^{-d} \right)$$ (equation 7.1)

Besides frequency and recency of usage of declarative facts, the context in which these facts occur also plays a role in the activation of these facts. This activation component will be called the spreading activation (Quillian, 1968), and represents the likelihood that one declarative fact will be needed if another one is currently being used. These likelihoods depend on the pattern of prior exposures with the declarative facts, as represented by the relatedness measure $R_{ji}$ between two facts $j$ and $i$ (Anderson & Lebiere, 1998; Anderson & Milson, 1989):

$$R_{ji} = \frac{F(W_j \& W_i)F(N)}{F(W_j)F(W_i)}$$ (equation 7.2)

where $F(W)$ are the frequencies of respectively fact $j$ and $i$, $F(N)$ the total number of exposures and finally $F(W_j \& W_i)$ is the number of co-occurrences of the facts $j$ and $i$. Equation 7.2 is sometimes referred to as associative strength (Anderson & Lebiere, 1998; Anderson & Milson, 1989), to indicate that the relatedness between two facts is determined by the environment. The model of declarative memory outlined here has been successfully deployed in predicting behavior in a variety of memory related cognitive tasks (e.g., Anderson et al., 1998; Anderson & Schooler, 1991; Van Rijn & Anderson, 2003).

**IMPLEMENTATION OF THE PERSONAL PUBLICATION ASSISTANT**

The Personal Publication Assistant is a personalization tool based on a personalized rational analysis of memory. Therefore the user models underlying the recommendations are constructed on an individual basis. In these models, each word that occurs in one of the abstracts of the user is represented by a combination of base-level activation (adapted from Petrov, 2006) and spreading activation from the other words in the model (Anderson &
Lebiere, 1998). These activation values can be calculated using the statistical properties of the words in the published abstracts of an individual researcher:

- The year in which it appears for the first time in one of the user’s abstracts,
- The year in which it most recently appears in one of the user’s abstracts,
- The frequency of appearance,
- The frequency of co-occurrence with another word.

Based on these properties, we create an individual representation of a researcher’s interests using the rational analysis described above. The Publication PA applies these individual user models to predict the relevance of words that occur in other scientific abstracts, by calculating how familiar these abstracts are. In the next sections, we will describe in more detail how the Publication PA calculates the base-level and spreading activation values, which words from the abstracts are taken into consideration, and how the system comes to a selection of the relevant information.

**THE RELEVANCE OF INDIVIDUAL WORDS IN THE USER MODEL**

With the equations that are provided by the rational analysis approach to declarative memory, we can calculate the base-level activation of a word based on its occurrences in publications of the user. The base-level activation can be seen as a measure of interest, with the most interesting words having the highest base-level activation.

For this application, an optimized version (Petrov, 2006) of the base-level equation discussed earlier (Equation 7.1) was used. In this equation (Equation 7.3), the decay parameter is fixed at .5 (and is reflected in Equation 7.3 as the square root operators) and a history parameter \( h \) is added:

\[
B = \ln \left( \frac{1}{\sqrt{t_1 + h}} + \frac{2n - 2}{\sqrt{t_1} + \sqrt{t_1 + h}} \right) \quad \text{with } h > 0
\]

(equation 7.3)

The first component of this equation reflects the most recent encounter of that word: the longer ago the word was encountered, the smaller the contribution is. The second component reflects the frequency of usage of the word. This optimized version of the base-level activation equation assumes that the encounters of the word are evenly spaced over time between the first encounter and the last encounter of the word. In the default equation, the base-level activation is a product of both recency and frequency. However, in a recommendation system, it might be useful to be able to change the balance between both factors. For example, a researcher might still be interested in work relating to older work, even though a recent project has resulted in a set of papers on a new topic. To enable this, we added the history parameter. The history parameter influences the effect of recency. Informally, a higher value for \( h \) spreads the publications over a longer time frame, decreasing the relative activation of a word that only recently came up in analyzed texts. In Experiment 1 we will demonstrate that the \( h \) parameter is an important parameter when recommending papers with the Publication PA.

**THE INFLUENCE OF CONTEXT ON WORD RELEVANCE**

Apart from the frequency and recency of usage of a word, the context in which a word occurs is also important. For instance, using the word model in your paper on user models should not elicit conference talks on fashion models. So, context words - like in this example user or rational - are important in determining the activation of words such as model or analysis. The context in which a word has occurred in previous abstracts is incorporated in the model by spreading activation (Equation 7.2), which reflects the personalized probability that a word
will be needed in connection with another word.

Recommendations occur by mediating the base-level activation of a word with the spreading activation of other words:

$$A_i = B_i + \sum_j WR_{ij}$$  

(equation 7.4)

In Equation 7.4, the base-level activation of the word \(i\) in a specific abstract is increased with the sum of all weighted connections with the words also found in that abstract. The connections are weighted because otherwise the ratio between the base-level activation and the spreading activation would be dependent on the number of associations. For this application the base-level activation of the connecting word \(j\) is used as the weight \((W)\), to scale down with the spreading activation from words that have a low base-level activation. This would be the case when the word \(i\) co-occurred often in the past with a word \(j\) that is present in the current abstract but which is not often used anymore (i.e., has a low base-level activation). This would cause the spreading activation to be high while the connection is less relevant at the current time, negatively influencing the selection of relevant papers.

FILTERING OF NON-CONTENT WORDS

The relatedness measure \(R_{ij}\) has shown to be a robust method of boosting the base-level activation as a function of the connectedness. That is, if two words always occur in tandem, the activation of the second word will be boosted when the first word is encountered. At the same time, a word that occurs in combination with many other words does spread less activation. In normal word usage, words as the and is spread only a small amount of activation because of this. In normal word usage, this effect makes sure non-content words do not influence base-level activations of other words too much. However, a problem might arise when the formulation of sentences in scientific abstracts differs from normal word usage. Because of spatial constraints, word usage in scientific abstracts might differ from normal written English. This might result in a lower frequency of function words, increasing their spreading activation (Equation 7.2), with a possibly negative influence on the eventual recommendations. To counter the unwanted influence of normally high-frequent words, these words are filtered from the data using a lexical database (Baayen, Piepenbrock, & Van Rijn, 1993). An analysis of the frequency distribution of words in both scientific abstracts and normal written English will demonstrate that filtering out high-frequent words will not interfere with how well an abstract represents the contents of a paper.

Analysis

To compare word usage in scientific abstracts with word usage in normal lexical content, the abstracts of all publications that appeared in the Cognitive Science Journal between 2004 and 2006 were used. Numeric symbols and punctuation were removed from the abstracts, resulting in a list of the words that were used in the abstracts. For each word, the frequency in all the abstracts was contrasted with an estimate of the normal frequency in written English, taken from the CELEX lexical database (Baayen, Piepenbrock, & Van Rijn, 1993). If a word was not found in the database because of spelling mistakes or terminology, the CELEX frequency was assumed 0, and the frequencies of multiple occurrences of a word were summed because in CELEX the frequencies of homonyms are counted separately. The CELEX frequencies were scaled to the total number of words of the abstracts to make them comparable.
Results

In Figure 7.1, the ratio between the CELEX word frequencies and the abstract word frequencies is plotted. We used a logarithmic scale for easier presentation. Figure 7.1 visualizes that the usage of words in scientific abstracts differs from the distribution of words used in normal written text. If the distributions were similar, then the dashed horizontal line would have represented the ratio. However, it becomes clear that a large part of the words used in the abstracts occur less often in normal written English; those are the words with a frequency ratio below one. Only a small part of the words occurs more often in normal written English. Thus, 2190 of the words used in the abstracts of the Cognitive Science Journal between 2004 and 2006 occur more frequently in scientific abstracts than in normal written English, while only 412 words occur more often in normal written English. However, those 412 words account for a large portion of the total amount of word occurrences found in the CELEX database (440,000 of the total of 740,000 occurrences of these words), while the 2190 words that are less frequent in normal written English generate less occurrences than the 412 high frequent words (300,000 of 740,000 word occurrences). This difference is caused by abstracts containing jargon and the tendency to use as little function words as possible, whereas in normal language these words are used very frequently. Thus, removing the words from the scientific abstracts that are most frequent in normal written English will not remove any of the important content words, as only words above the dashed line in Figure 7.1 are deleted, while words below the dashed line in Figure 7.1 are the words that are relevant to the Publication PA.

Selection of Relevant Abstracts

The final part in the recommendation is finding the amount of activation for each paper and presenting the user with a ranking or selection. In general, abstracts in which many words have a high activation, have a high degree of familiarity to the researcher, and are thus interesting enough to select. To compare the relevance of papers with each other, every abstract has to be represented by a single value. One solution would be to sum the activations of all the words in the conference abstract. However, simply summing activation values would result in a bias towards longer abstracts. To counteract this bias, we chose to average the activation of the words that occur in both the abstract and the user model. This means that the effect of abstract length is neutralized, while still taking all activation values of the words in the abstracts into account.

Experiments

To validate the Publication PA, we first analyzed what the influence of the $h$ parameter
is. Second, we performed a user study with a sample of researchers from the field of cognitive science, asking them to rate how much a recommended abstract aligned with their interests.

**EXPERIMENT 1: HISTORY PARAMETER ANALYSIS**

**Methods**

We analyzed the behavior of the Publication PA with four different values for the history parameter: \( h = 0.0001 \); \( h = 0.1 \); \( h = 10 \); \( h = 1000 \). The parameter values were chosen to maximize a potential effect. The only other parameter in the system (the decay parameter \( d \)) was left at the default value of 0.5.

As a test set, we took the abstracts of the publications of professor John R. Anderson, for as far as indexed by PsycINFO\(^{15}\). When visually inspecting his publication record, it shows some stable interests over time, but also some changes in interest. As a cognitive modeler, almost all of Anderson’s publications deal with cognition and the cognitive architecture he developed, ACT-R. However, a change in focus can be observed. From the start of his career, Anderson’s interests seem to be related to learning and memory (as witnessed by for instance Anderson & Bower, 1972, 1973), whereas more recently he seems to have developed an interest in functional brain imaging techniques (e.g., Anderson, 2007b; Anderson, Albert, & Fincham, 2005). These trends should also be visible if we apply different parameter values to the \( h \) parameter and construct different user models.

**Results**

To compare the user models that were constructed with the various values for the \( h \) parameter, we ordered the words in the user models according to their activation values. Thus, the ordering represented the estimated importance of a word for a person’s interest. Figure 7.2 presents the rank order values of various words that are exemplary of the trends found in Anderson’s publication record. Small values of \( h \) indicate that the relative influence of more recent publications increases; this effect is reflected by the decreasing rank (and thus increasing importance) of the words functional and imaging for decreasing values of \( h \). These words do all relate to the recent research interests. On the other hand, the words memory and experiments show the opposite trend. This reflects a shift of interest from proto-typical memory-related research in which multiple experiments are presented per paper. Also, the ranks of some words stay constant with changing \( h \) values. ACT-R and cognitive are words that appear in both recent and past abstracts of professor Anderson, indicating a stable interest in these concepts.

This qualitative inspection of the results leads us to believe that the history parameter plays an important role in the selection of relevant abstracts, because it determines the ranking of the activation values. What the optimal setting for this parameter should be might be determined in a large user study in which we ask participants to rate the relevance of abstracts that are selected using various values for the history parameter (as has been done for this analysis). However, given the personal nature of interest, it seems better to leave the optimal setting to the user, for example, by presenting the user with the possibility to set this parameter in the user interface. To evaluate the performance of the Publication PA independent of the relative importance of word usage history, we decided to run the user study with \( h \) set to 10.
Experiment 2: User Study

We performed a user study to evaluate the recommendations provided by our abstract recommender system. We asked 10 researchers (2 full professors, 2 associate professors, 5 assistant professors, and 1 post-doc) from various subfields of cognitive science and from various countries how much they are interested in a paper after reading the abstract.

Methods

For each of the researchers, we constructed user models based on the abstracts of their published work insofar it was indexed by PsycINFO. Next, we ordered all abstracts from the last three volumes (2004-2006) of the *Cognitive Science Journal* according to their relevance for an individual researcher, based on the researcher’s published abstracts.

From the ordered list of abstracts, we presented the top five abstracts, the bottom five abstracts (that is, the least relevant abstracts), and five abstracts from the middle of the list to each researcher. The presentation order of these 15 abstracts was randomized, to eliminate any effects from expectations about the presentation order. We asked the researchers to indicate with a grade between 0 and 9 how much they are interested in the papers, based on the abstracts. We adopted this scale from similar work done by Dumais and Nielsen (1992) in order to be able to make a comparison between their approach and ours. Following Dumais and Nielsen (1992), we characterized the meaning of the rates as follows:

- 8-9: right up my alley
- 6-7: good match
- 4-5: somewhat relevant
- 2-3: I’m following it, sort of
- 0-1: how did I get this one?

Results

To analyze the performance of the Publication PA, we applied two measures of relevance:

- Mean rated relevance,
- Precision.

The precision and mean rated relevance were applied to each of the three groups (top 5, middle 5, bottom 5). Because it is not feasible for the participants to rate all available abstracts from the *Cognitive Science Journal* between 2004 and 2006 (129 abstracts), we did not calculate the rate of recall, as is often used in these kinds of applications (Salton & McGill, 1983). However, the recall rate is implicitly accounted for in the measures we did apply.
Mean rated relevance

We analyzed the relevance rates given to the abstracts for each group. Figure 7.3 shows the means of the rates per group. Welch t-tests between the groups reveal that the rates given for the top 5 abstracts differ significantly from the other two groups (t=4.20, df=86.54, p<0.001 for the top 5 vs. the bottom 5 and t=3.64, df=94.06, p<0.001 for the top 5 vs. the middle 5). The rates for the bottom five abstracts did not differ significantly from the rates for the middle five abstracts. This is in line with the observation that in multidisciplinary journals such as Cognitive Science, the relevance rate does not decrease linearly, but instead that only a small part of the published papers is relevant for a researcher, and the rest is not.

If the Publication PA would not be able to suggest relevant papers, this would mean that in all three groups the number of highly rated papers would be equal on average. However, if this were the case, we would not be able to observe significant differences in the mean rated relevancies between the top 5 recommended papers and the other two groups. The fact that we do find this difference indicates that the system is able to provide a meaningful rank order in which the higher rated papers will be ranked higher.

Precision

Precision of retrieval is usually defined as the number of relevant documents that is retrieved relative to the total number of documents retrieved (Salton & McGill, 1983). Following the meanings of the anchor points of the scale we provided to the participants, relevance should be taken as rated with 4 or higher. Using Equation 7.5, the precision of the Publication PA in the top 5 recommended abstracts is p = 0.58.

\[
    p = \frac{|\{\text{rating} > X\} \cap \{\text{retrieved abstracts}\}|}{|\{\text{retrieved abstracts}\}|}
\]

(equation 7.5)

Because this notion of relevance may be considered arbitrary, we also calculated the precision of the Publication PA with different assumptions on relevance. For example, we calculated precision under the assumption that only abstracts rated 8 or higher where relevant, or that all abstracts rated 2 or higher were relevant. In Figure 7.4, the results of this analysis are presented. The figure shows that, although precision declines with a more stringent notion of relevance, the precision in the top 5 recommended abstracts is always higher than in the other two groups.

DISCUSSION AND CONCLUSION

With our experiments, we demonstrated both the flexibility of the Publication PA and its applicability. With only one parameter, we could change the recommendations of the system in such a way that the relative influence of older papers changed, resulting in different recommendations.

With the h parameter at a fixed value, we demonstrated that the Publication PA can
provide meaningful recommendations for individual users. Two observations from this experiment should be further discussed.

From both the precision measure and the mean ratings, it becomes clear that there is no real difference between the group of abstracts from the bottom of the order list of abstracts from *Cognitive Science Journal* (2004-2006) and the ‘middle’ group. This shows that from a large collection of papers, only a very small subset is relevant for a particular user, underlining the need for filtering mechanisms or recommender systems.

Figure 7.4 shows that the mean rated relevance for the top 5 recommended abstracts is 4.5. This qualifies as *somewhat relevant*, but not *right up my alley*. We contribute this to the nature of the data set we used to recommend abstracts from. *Cognitive Science Journal* is a highly multidisciplinary journal, accepting papers from a wide range of research areas (as witnessed for instance by the set of keywords authors can use when submitting, published on the website of the Cognitive Science Society16). As a result, papers addressing very specific topics, that may be *right up my alley*, will be presented to other, more specialized, journals. Thus, the ratings provided by our participants might be a bit lower than expected, because abstracts that would be rated as *right up my alley* were probably underrepresented in the data set.

**RELATED WORK**

The problem of matching researchers and papers has been addressed before, in the context of systems that use Latent Semantic Indexing (*LSI*) (Dumais, 2003; Dumais & Nielsen, 1992; Foltz & Dumais, 1992). Our approach deviates from these earlier attempts in a number of ways. *LSI* assumes that the similarity of two documents is reflected by the similar word frequency distributions that are manifest in these documents (Deerwester et al., 1990; Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998). However, instead of taking the raw frequency statistics into account, *LSI* performs a mathematical analysis (singular value decomposition) that is capable of higher-order inference. That is, *LSI* calculates the probability of each of the words occurring in a document, given multiple documents.

Instead of *LSI*, the measure of semantic relatedness that we apply, associative strength (Anderson & Lebiere, 1998; Anderson & Milson, 1989), is equivalent to Point-wise Mutual Information (*PMI*), given a reasonably large data set (Farahat, Pirolli, & Markova, 2004). *PMI* is also based on the statistical properties of the documents, but, in contrast to *LSI*, *PMI* is a direct measure of the likelihood that one word will occur, given the presence of another. As a measure of *semantic similarity*, *PMI* has been shown to perform equal to or better than *LSI* (Turney, 2001). We expect therefore that when *PMI* will also be a better representation of semantic relatedness than *LSI*.

Besides the method of calculating the semantic relatedness, also the corpus of text on which it is performed differs. Dumais and colleagues (Dumais & Nielsen, 1992; Foltz & Dumais,
used a fixed semantic space for all users of their system. Recently, however, it has been shown that the choice of corpus greatly influences the semantic distance, even when applying the same measure of semantic relatedness (Lindsey, Veksler, Grintsvayg, & Gray, 2007). By contrast, we constructed personalized semantic spaces for individual users. That is, the associations between words in the semantic space reflect the semantic relatedness as apparent from the statistical properties of word usage in the abstracts of one user. This obviously will result in more individualized recommendations, because only the associations between words that a single researcher would also make, are present. Also, the problem of corpus selection does not arise, because the corpus used is already the best possible representation of a researcher’s interest, namely her own publication record.

When it comes to the performance of the Publication PA as compared to the approach taken by Dumais and Nielsen (1992), the Publication PA seems to perform equally well. Dumais and Nielsen (1992) report a precision of $p = 0.51$, slightly lower than our value of $p = 0.58$. However, the computation of precision differs between the two approaches. In general, comparison is difficult because of the different nature of the data sets used. While the abstracts from the *Cognitive Science Journal* are very multi-disciplinary and thus very diverse, the abstracts used by Dumais and Nielsen (1992) are from a very specialized conference (*ACM Hypertext’91*). This difference in diversity of topics included in the data sets could explain the difference in mean rated relevance between the Publication PA (4.5) and the system by Dumais and Nielsen (5.75). In the Dumais and Nielsen experiment, both the abstracts and the researchers they are being assigned to, are specialized in hypertext. Therefore, the mean relevance of the data set for the researchers is already higher than in our experiment.

To a certain extent, our work bears resemblance to the work of Pirolli and colleagues towards Information Foraging (Fu & Pirolli, 2007; Pirolli, 2005; Pirolli & Card, 1999; Pirolli & Fu, 2003). They provided a rational analysis of how users search for relevant information, and applied this to information search on the World Wide Web. This way, they were able to model web navigation aspects of a typical user. In Information Foraging theory, the likelihood that a certain document or webpage is relevant is based on the base-level activation of the words in that document and the spreading activation from the words in that document to the words in the search query. Similarly, the Publication PA computes the relevance of a paper based on the base-level activation of the words in the abstract and the spreading activation from the words in that abstract to the words in the user model. One of the important components of the Publication PA is the construction of the user model, which ensures that only words that are relevant for an individual researcher are considered in computing the relevance of an abstract.

However, the Information Foraging models differ in that they capture the information search behavior of a typical human being serving the web, whereas the Publication PA is a personalization tool, and is intended to model the information needs of an individual researcher. As outlined above, the semantic relatedness estimates applied by the Publication PA are therefore personalized for each individual researcher, resulting in different behavior of the model for each researcher.

CONCLUSION

In this paper, we proposed a method for the personalization of information selection, based on rational analysis and cognitive architectures. We developed an application, the Personal Publication Assistant (Publication PA), for the recommendation of relevant scientific abstracts to researchers, based on their publication record to date. In two experiments, we
analyzed the behavior of the Publication PA and found that it is a flexible and adaptable system, as well as an adaptive system. From Experiment 1 we concluded that users of the Publication PA can adapt the nature of the recommendations to their own personal wishes, using only one parameter. In a final version of the interface, this parameter could be controlled by a slider bar. Some researchers might be only interested in their current topic, for instance because they have just switched research topics. They can choose a low value for this \( h \) parameter. Researchers that would rather want to follow what is being published in research fields they previously published in may choose a high value of the \( h \) parameter.

Experiment 2 demonstrated that the Publication PA can select relevant papers for individual researchers. Papers that were recommended by the system were rated higher by the participants than papers that were not recommended.

The techniques applied in the Publication PA might also be applied to develop recommender systems in other domains in which personalized information retrieval is desirable. The domain should be primarily characterized by textual information sources, such as conference or journal papers, and the users should also be characterized by textual testimonials of their interests. Two examples of the wider applicability of the method of information selection that we proposed here are the problem of assigning manuscripts submitted to a conference to reviewers, and the problem of selecting relevant press bulletins from the stream of bulletins provided by press agencies world wide. We will discuss both these examples and hint at an implementation of our technique.

The assignment of manuscripts submitted to a conference to reviewers is a problem very similar to the selection of relevant abstracts for a reviewer. Even though reviewers can often indicate their areas of expertise, it is hard for conference program chairs to match every submission to the most qualified reviewers. Since the area of expertise of a reviewer is reflected in his or her publication record, user profiles that reflect the areas of expertise could be generated based on the publication record. By matching the profiles against each submitted abstract, the best-suited reviewer for each abstract will be associated with the highest relevance score. This way, conference chairs can easily assign submitted manuscripts to reviewers without having to rely on the reviewer’s own opinion of his or her expertise, or without having to burden them with long questionnaires about their fields of research.

Press agencies produce many bulletins a day, often over 12,500 bulletins a year.\(^\text{17}\) A reporter trying to read the most important press bulletins for his or her interests has to make a selection from this vast amount of information. Although press agencies often tag their bulletins or assign them to a certain category, it is easy to miss the one that is important. By creating profiles of reporters based on the news articles they have written over the years, an application similar to the Publication PA could make a meaningful selection for them.

A cognitive model of information selection can thus guide the development of a recommender system, because it provides insights in which features from the pieces of information are relevant for the selection process. The analysis suggests that the features that people that are engaged in retrieving relevant information use are the history of usage of words, and the co-occurrence of words. By incorporating these features in the same way as a cognitive model of human memory does, we have created a successful Publication PA, that for example can decrease the work load of individual researchers attending a conference by creating a preselection in the conference proceedings.
A Comparison between Decision Making and Memory Models for Literature Selection


LITERATURE SELECTION

In 2006, the number of scientific publications in the relatively small ISI subject category Information Science & Library Science was 2054. In other words, researchers working in this area had to scan through over 2,000 papers a year to keep up with the current developments. However, this number is, if anything, an underestimation of the total number of potentially relevant papers, as this number only holds if a researcher is interested in a single subject area. In practice, most researchers work on the intersection of multiple domains, increasing the number of potentially relevant papers enormously. Not only professionals in the scientific domain are confronted with masses of potentially relevant information. Also, government or business employees often need to decide which of numerous reports, leaflets, and bulletins to read, and which to ignore - a challenge that is aggravated by the continuously increasing amount of information that is available online. For instance, many press agencies produce over 12,500 bulletins a year. Reporters trying to read the most important ones have to make selections, and although the agencies often tag their bulletins, the sheer mass of information makes that it is easy to miss important ones.

In this paper, we will focus on one solution to this problem: recommender systems. Typically, corresponding decision aids automatically come up with a pre-selection of information that is worth further consideration, saving institutions, firms, and people parts of the time and effort otherwise required to separate the relevant from the irrelevant. In particular, here we will evaluate six models that can solve the problem of information selection for the scientific domain.

All models select relevant scientific papers from a large collection of scientific abstracts. They include (i) the Publication Assistant (Van Maanen et al., in press), a recommender system that was recently developed to assist scientists in identifying relevant articles. We will compare the performance of this system to that of (ii-iv) three simple decision heuristics, including a unit-weight linear model (see Dawes, 1979; Dawes & Corrigan, 1974; Gigerenzer & Goldstein, 1996), and two lexicographic rules, called take-the-best (Gigerenzer & Goldstein, 1996), and naiveLex. We will also pit all models against (v-vi) two more complex linear weighted additive models, one being Franklin’s rule (Gigerenzer & Goldstein, 1999) and the other multiple regression (Slovic & Lichtenstein, 1971).

While we do not aim to model the cognitive processes that are actually going on when scientists make literature choices, except for multiple regression all models tested here are grounded in cognitive theories. The Publication Assistant is a memory model that is based on the rational analysis framework (Anderson, 1990; Oaksford & Chater, 1998), as incorporated in the ACT-R cognitive architecture (Anderson, 2007a). The heuristics are models of decision making that are grounded in the fast and frugal heuristics framework (Gigerenzer, Todd, &
the ABC Group, 1999). The linear weighted additive model, Franklin's rule, is also a model of decision making (Gigerenzer & Goldstein, 1999). All models are common in the memory and judgment and/or decision making literature.

In what follows, we will give an outline of the Publication Assistant. Next, we will introduce the five alternative models. In an experiment, we will then evaluate the models’ performance in predicting scientists’ literature preferences.

THE PUBLICATION ASSISTANT: A MEMORY MODEL

An example of the problem addressed in this paper is the selection of relevant talks when attending a large, multi-track scientific conference such as the Annual Cognitive Science Conference. The information selection process starts when a researcher registers and receives a copy of the conference program. For instance, a strategy often employed by many conference attendees is to scan talk titles, author names, or abstracts for words or names that sound familiar. If an entry contains enough interesting words, it is selected for more careful reading, and the corresponding talk might be attended. In order to determine if a word qualifies as interesting in the context of the conference, a researcher might assess whether she has used the word in her own research in the past. The assumption is that the words used by someone in the context of their own research reflect their scientific interests. The Publication Assistant is a literature selection tool that could be run over a (digitized) conference program prior to attending the conference. The model recommends talks a given scientist might find useful to attend, saving that researcher the time and effort required to scan the conference program on his own. To this end, the model searches through the scientist’s own work, examining in how far words that appear in conference abstracts also occur in the scientist’s work. Specifically, the model bases its recommendations on the following properties an abstract’s words:

**Recency of occurrence in the scientist’s own work**
- The year in which a word from a conference abstract appears for the first time in the abstracts the scientist has published in the past,
- The year in which a word from a conference abstract appears for the last time in the abstracts the scientist has published in the past,

**Frequency of occurrence in the scientist’s own work**
- The frequency of appearance of a word from a conference abstract in the abstracts the scientist has published in the past,
- The frequency of co-occurrence of a word from the conference abstract with another word in the abstracts the scientist has published in the past.

Based on these properties, the model creates an individual representation of a researcher’s interests. The Publication Assistant applies these user models to predict the relevance of words that occur in other scientific abstracts, essentially estimating how familiar the contents of these abstracts would be to the scientist. In the next section, we will describe in more detail how the Publication Assistant estimates familiarity.

**MODEL EQUATIONS**

The Publication Assistant works like a model of the contents of a researcher’s memory. Its equations are based on Anderson and Schooler’s (1991) rational analysis of memory. According
to their analysis, the probability that a fact (e.g., a word) stored in memory will be needed to achieve a processing goal can be predicted from the organism’s pattern of prior exposure to the corresponding piece of information. For example, the probability that a fact about a scientific topic is of relevance to a researcher may depend on the frequency and recency of his writings about it in the past. Frequency and recency, in turn, feed into a memory currency called base-level activation, which influences a researcher’s familiarity with the fact. These relations are captured by Equation 8.1, in which \( B \) stands for the base-level activation of a fact \( i \), \( t_i \) stands for the time that has passed since the last exposure to that fact, and \( d \) represents the speed with which the influence of each exposure decays away. The summation takes place over all \( n \) previous encounters with the fact.

\[
B = \ln \left( \sum_{i=1}^{n} \frac{t_i^{-d}}{t_i + h} \right) \quad \text{(equation 8.1)}
\]

Besides frequency and recency of encounters with facts, the context in which these facts occur also plays a role in the activation of the facts. This spreading activation (Quillian, 1968) component represents the likelihood that a fact will be needed if another one is currently being used. These likelihoods depend on the pattern of prior exposures with the facts, as represented by the relatedness measure \( R_{ji} \) between two facts \( j \) and \( i \) (Anderson & Lebiere, 1998; Anderson & Milson, 1989):

\[
R_{ji} = \frac{F(W_j \& W_i)}{F(W_j)F(W_i)}F(N)
\]

\[
\text{(equation 8.2)}
\]

where \( F(W_j) \) and \( F(W_i) \) are the respective frequencies of facts \( j \) and \( i \), \( F(N) \) is the total number of exposures, and \( F(W_j \& W_i) \) is the number of co-occurrences of the facts \( j \) and \( i \).

With the equations that are provided by the rational analysis of memory, one can calculate the base-level activation of a word based on its occurrences in publications of the user. However, rather than using Equation 8.1 directly, the Publication Assistant uses Petrov’s (2006) version of it. In Equation 8.3, the decay parameter is fixed at .5 and a history factor \( h \) is added, which represents a free parameter:

\[
B = \ln \left( \frac{1}{\sqrt{t_i + h}} + \frac{2n - 2}{\sqrt{t_i + h} + \sqrt{t_i + h}} \right) \quad \text{with } h > 0
\]

\[
\text{(equation 8.3)}
\]

To stick to the example of selecting abstracts from a conference program, the Publication Assistant makes recommendations by combining the base-level activation of a word \( i \) with the weighted base-level activation of related words \( j \) in the abstract (Pirrolli & Card, 1999):

\[
A_i = B_i + \sum_j B_j R_{ji}
\]

\[
\text{(equation 8.4)}
\]

To compare the relevance of abstracts with each other, each one is represented by the average activation of the words that occur in it. In a comparison of two abstracts, the Publication Assistant then recommends the more activated one. Abstracts in which many words have high base-level and spreading activation values have a high match with the researchers own word usage, and thus may be more interesting. The Publication Assistant’s recommendations are thus based on the structure of the environment of a particular researcher. In particular, the structure of word usage in previously published abstracts. The only parameter that may be varied is the history parameter \( h \), which represents the relative importance of recency versus frequency in determining activation. In the research reported here, we kept \( h \) constant at the same value reported in Van Maanen et al. (in press).
To evaluate the performance of the Publication Assistant in predicting scientists’ literature preferences, we compared it to five alternative models. While the Publication Assistant essentially mimics a model of memory, these alternative models have originally been proposed as decision strategies in the judgment and decision making literature.

In particular, we focus on a class of models that have been prominent in the fast and frugal heuristics framework. According to this framework, humans (and other organisms) often make decisions under the constraints of limited information processing capacity, knowledge, and time – be they about the relevance of scientific articles, or the likely performance of stocks, or the nutritional value of food. Such decisions can nevertheless be made successfully because humans can rely on a large repertoire of simple decision strategies, called heuristics. These rules of thumb can perform well even under the above-mentioned constraints. They do so by exploiting the structure of information in the environment in which a decision maker acts and by building on the ways evolved cognitive capacities work, such as the speed with which the human memory system retrieves information.

One of the heuristics tested here, the unit-weight linear model, is particularly simple, requiring no free parameters to be fitted. Related models have proved to be almost as successful (or even better) in predicting unknown events and quantities as multiple regressions (see Dawes & Corrigan, 1974; Dawes, 1979). Just as the unit-weight linear model, also naiveLex dispenses with all free parameters. If these two particularly simple heuristics predicted scientist’s literature preferences successfully, then they would simplify the selection of abstracts more than the Publication Assistant does. In order to be considered a useful tool, the Publication Assistant should thus be able to outperform these models. Take-the-best is a little more complex, requiring one free parameter to be fitted for each individual scientist. Take-the-best and related models have been found to be, on average, more accurate than multiple regression in predicting various economic, demographic, and environmental, variables (e.g., Czerlinski, Gigerenzer, & Goldstein, 1999). Finally, the most complex models tested here, Franklin’s rule and multiple regression, require for each individual researcher as many free parameters to be fitted as there are words in the abstracts under consideration. While these two models are prominent in the judgment and decision making literature, due to their large complexity they are not considered heuristic decision strategies in the fast and frugal heuristics framework. Rather, they are often used as benchmark to evaluate the performance of heuristics in model comparisons (Czerlinski, Gigerenzer, & Goldstein, 1999; Gigerenzer & Goldstein, 1996).

Lexicographic Heuristics: Take-the-Best, NaiveLex

The first model to be considered here is take-the-best. To make literature recommendations, take-the-best uses attributes of articles as cues. In our context, cues are the words that occur in an abstract. If a word also occurs in a scientist’s own publication, then it is considered a positive cue, suggesting that an abstract is of interest to that scientist. Take-the-best considers all cues sequentially (i.e., one at a time; hence lexicographic) in the order of their validity. The validity of a cue is the probability that an alternative A (e.g., an article) has a higher value on a criterion (e.g., relevance for a researcher) than alternative B, given that alternative A has a positive value on that cue and alternative B does not. In a comparison of two abstracts, take-the-best bases a decision on the first cue that discriminates between the abstracts, that is, on the first cue for which one abstract has a positive value and
the other one does not. The heuristic is defined in terms of three rules:

1. Look up cues in the order of their validity.
2. Stop when the first cue is found that discriminates between the abstracts.
3. Choose the abstract that this cue favors.

The second lexicographic model, here called naiveLex, is identical to take-the-best, except that it does not estimate the validity order of cues. Rather, cues are simply considered in the order of the frequency of occurrence of the corresponding words in each researcher’s published abstracts. This aspect of the model is similar to the Publication Assistant, in which the word frequency is also taken into account (but weighted with recency).

A UNIT-WEIGHT-LINEAR HEURISTIC

Lexicographic heuristics such as take-the-best can avoid going through all words (i.e., cues) from an abstract, which can save effort, time, and computations once the order of cues is known. Unit-weight linear heuristics, in contrast, integrate all cues into a judgment by adding them. These models can nevertheless simplify the task by weighing each cue equally (hence unit-weight). In a comparison of two abstracts, it reads as follows:

1. For each abstract, compute the sum of positive cues.
2. Decide for the abstract that is favored by a larger sum.

WEIGHTED-ADDITIONAL MODELS: FRANKLIN’S RULE AND MULTIPLE REGRESSION

Franklin’s rule (Gigerenzer & Goldstein, 1999) is similar to the unit-weight linear heuristic, but instead weights all the cues by their validities prior to summation. (The cue validities are identical to those relied on by take-the-best.) Multiple regression, in turn, estimates the weights of the cues by minimizing the error in the calibration set using maximum likelihood estimation. In a comparison of two abstracts, Franklin’s rule and multiple regression read as follows:

1. For each abstract, compute the weighted sum of cues.
2. Decide for the abstract that is favored by a larger sum.

EXPERIMENT

To compare the Publication Assistant to the alternative models’ capability of predicting actual scientist’s literature preferences, we re-analyzed data from a study by Van Maanen et al. (in press, Experiment 2). They had asked researchers from the field of cognitive science to rate how much they were interested in a paper after reading the abstract. In this study, Van Maanen et al. had found that the Publication Assistant could fit researcher’s interests reasonably well; however, they did not compare its performance to that of alternative models.

METHODS

Participants

Ten researchers (2 full professors, 2 associate professors, 5 assistant professors, and 1 post-doc) from various subfields of cognitive science and from various countries were asked to participate.

Procedure

For each of the researchers, Van Maanen et al. (in press) constructed user models of the Publication Assistant based on the abstracts of their published work insofar it was indexed by PsycINFO. They then ordered all abstracts from the last three volumes (2004-2006) of the
Cognitive Science Journal according to the predicted relevance for the researcher, based on the researcher’s published abstracts.

From the ordered list of abstracts, they presented each researcher the top five abstracts, the bottom five abstracts, and five abstracts from the middle of the list. For each researcher, the presentation order of these 15 abstracts was randomized. Each researcher indicated with a grade between 0 and 9 how much he or she was interested in the papers, based on the abstracts.

Analyses

To compare the performance of the Publication Assistant to that of the five alternative models in predicting each researcher’s ratings, we ran a cross-validation. To this end, we constructed paired comparisons of all 15 abstracts for each participant individually (210 pairs). We divided each participant’s abstracts pairs randomly into two parts. The first part represented the calibration set in which we calculated for each participant that person’s optimal values for the free parameters in take-the-best, Franklin’s rule, and multiple regression, respectively. That is, we identified the parameter value at which each model would correctly predict the largest proportion of literature preferences. Take-the-best, Franklin’s rule, and multiple regression will therefore be referred to as the calibrated models.

We used these optimal values to compute the proportion of preferences consistent with each model in the other half, the validation set, where the models’ generalizability is evaluated. For each partition, we also computed the proportion of preferences consistent with the three not-calibrated models (the Publication Assistant, naiveLex, and the unit-weight linear heuristic. The free parameter of the Publication Assistant, $h$, we set to 10. In fitting the very same participants as we do here, van Maanen et al. (in press), had found this value to work reasonably well. The other two models were parameter free in this respect.

We ran these analyses for a subset of possible sizes of the calibration and validation sets; that is, we first computed the proportion of each model’s correct predictions for a calibration set size of 1 and a test set size of 209, then for a calibration set size of 11, and a test set size of 199, and so on. The larger the size of the calibration sets, the larger is the sample of paired comparisons from which the parameterized decision models can estimate an individual researcher’s interests, that is, the more “experience” these models can accumulate before making their predictions. This procedure was repeated enough times to average out noise due to the random selection of calibration sets.

Figure 8.1 The performance of the non-calibrated models. A-J represent individual participants. PA: Publication Assistant; UWL: unit-weight linear heuristic.
RESULTS

When comparing the Publication Assistant with the other non-calibrated models (naiveLex and the unit-weight linear model), we found that the three models performed differently for different participants (Figure 8.1). The Publication Assistant made the most correct inferences for three participants (A, D, and J), while unit-weight linear heuristic scored outperformed the other two non-calibrated competitors on four occasions (B, C, E, and H). NaiveLex scored best for three participants (F, G, and I). Overall, the performance of the models did not differ much (mean_pA=0.60, mean_naiveLex=0.59, mean_UWL=0.60).

For each of the 10 participants, Figure 8.2 shows the proportion of correctly predicted preferences for the three calibrated models as a function of the size of the calibration set. As one would expect, for all participants the accuracy of the predictions of the parameterized models increases with the size of the calibration set. Of the calibrated models, Franklin’s rule was consequently outperformed by the take-the-best heuristic and the multiple regression model, which performed equally well, but differed among participants. Take-the-best was the best predictor for participants B, C, E, G, I, and J, while the regression model performed best for participants A, D, F, and H. Overall, take-the-best performed best (mean_TTB=0.84, mean_MR=0.81, mean_Franklin=0.71).

DISCUSSION

We examined the ability of six models to predict scientists’ literature preferences: (i) the Publication Assistant, a recommender system that is based on a rational analysis of memory and the ACT-R architecture; (ii-iv) three simple heuristics, including take-the-best, a naive lexicographic model, and a unit-weight linear model, and (v-vi) two complex weighted-additive models, Franklin’s rule and multiple regression.

For some participants and calibration set sizes, the regression model outperformed take-the-best. One reason why take-the-best did not fare as well as multiple regression on every occasion might be that the structure of information available in the abstracts was not well suited for this simple heuristic (Martignon & Hoffrage, 2002). For instance, take-the-best essentially bets on a noncompensatory information structure, always preferring the most
valid discriminating cue to all others. In the domain of literature selection, such information structures might not be prevalent. To give an example, the words “Memory” and “Retrieval” might be equally good predictors of some cognitive scientist’s research interests.

One result was that the performance of the non-calibrated models differed between participants. However, it should be realized that naiveLex and the Publication Assistant only differ with respect to the use of the recency component. Both models use the frequency of words in published abstracts in the same way. Therefore, the difference in the participants in which naiveLex is the better recommender may be attributed to the importance that these participants contribute to recency. That is, the Publication Assistant overestimates the importance of more frequent words in the published abstracts. Thus, recommendations of the Publication Assistant could improve if we would allow the $h$ parameter to be fit individually.

The fact that the non-calibrated models performed differently between participants is in agreement with other findings in the judgment and decision making literature, where large individual variation in people’s use of decision strategies are commonly observed (Bröder & Gaissmaier, 2007; Mata, Schooler, & Rieskamp, 2007; Pachur, Bröder, & Marewski, 2008). In this respect, it is somewhat surprising that take-the-best consistently outperforms all other competitors. Based on the literature, we would have expected that the most useful approach designing recommender systems would have been to build different systems for different users, depending on which model predicts the respective scientist’s preferences best.

**WHY WAS THE PUBLICATION ASSISTANT OUTPERFORMED BY THE CALIBRATED MODELS?**

Take-the-best, Franklin’s rule, and the regression model learned about the scientists’ interests directly from the paired comparisons between abstracts that were included in the calibration sets. The Publication Assistant, in turn, was trained on a participant’s published abstracts (Van Maanen et al. in press), under the assumption that word frequencies in those abstract would reflect the participants’ interests. While this way of training the model better reflects real-life situations of information selection, in which people’s appraisal of items (such as abstracts) is often unknown, it might have been detrimental for the model’s performance. In addition, our results complement findings by Lee, Loughlin, and Lundberg (2002), who, in a study on literature search, examined the performance of a simple heuristic in identifying articles that are relevant to a given topic of interest (e.g., eyewitness testimony). Their analyses show that a researcher going by a variant of take-the-best would have had to search through fewer articles in order to find the relevant ones than a person behaving in accordance with a weighted-additive model.

**CONCLUSION**

In this paper, we evaluated the ability of cognitive models of memory and decision making to serve as literature recommender systems. Three of the models were trained on the participants’ published abstracts (the non-calibrated models), while three other models were allowed to train on a calibration set that contained abstracts that were also part of the paired comparisons in the validation set. This second type of training generally yields better results, but at a cost of a less realistic training situation. The non-calibrated models showed large individual variability, suggesting that for successful recommendation, the best predicting model should be determined first. Future work will show if this result is generalizable to other domains of literature recommendation and information search.

To conclude, in today’s world of mass media, the choice which information to attend to,
and which to ignore becomes an ever more important challenge for professionals. Automatic recommender system might help to cope with these demands of the information age - savings in time and effort that can eventually be invested elsewhere. We hope that comparisons between different approaches, such as the ones tested here, help along that way.