2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

Impacts of Analysts' Cognitive Styles on the Analytic Process

Eugene Santos, Jr.*, Hien Nguyen+, Fei Yu*, Deqing Li*, John T. Wilkinson*,

*Dartmouth College
Thayer School of Engineering
8000 Maclean
Hanover, NH 03755
Eugene.Santos.Jr@Dartmouth.EDU

+University of Wisconsin-Whitewater
Dept. of Math and Computer Science
800 W. Main Street
Whitewater, WI 53190
nguyenh@uww.edu

Abstract

A user's cognitive style has been found to affect how they search for information, how they analyze the information, and how they make decisions in an analytical process. In this paper, we propose an approach that uses Hidden Markov Models (HMM) to dynamically capture a user's cognitive style by automatically exploring the sequence of actions and relevant information with respect to the content of the actions. The evaluation results show that our HMM model achieves an average of 72% recall with the APEX 07 collection. We also study the link between a user's cognitive style and the various attributes relating to document content during an analytical process. The results show that the "analytic" group tends to focus on documents with significantly more specific information than the "wholist" group. The specific/general attribute of documents can help us in classifying a user's cognitive styles automatically.

1. Introduction

With the increasing availability of online resources, collecting information on the Web and analyzing data play important roles in today's problem solving task. It has be found that a user's cognitive styles affect their searching/browsing behaviors, their assessment of the relevancy of a web page, and their decision making processes (for example: [5],[23],[6],[11]). However, little research has been conducted that explore the impacts of a user's cognitive styles on an analytical process. The challenges here are three fold: First, popular information retrieval and filtering systems on the Web do not take into account a user's cognitive styles even though this factor is known to affect a user's information seeking behaviors [6], and a user's assessment of text summarization [11]. Secondly, the unavailability of well-defined and relevant testbeds poses critical challenges to the process of capturing the goal-oriented and compound nature of an analytic process. Third, the open and diverse nature of the Web creates uncontrollable noise/influences such as environmental factors (variations in interfaces, dynamics of web information, etc.) or even credibility of participants that can affect the results of such a study. Therefore, in this paper, we explore the problem of the impacts of a user's cognitive styles on analytical processes in the domain of intelligence analysis. Our results can also be used and extended to the Web community to solve analytical problems done in Web settings.

Given a particular analysis task, it is likely that different people will take different approaches to solving the task. For example, political analysts will vary in how they perform their assessments of tasks such as "What will be the political repercussions of passing health care reform legislation?" Some will delve very deeply into the details of how individual politicians will be affected in upcoming elections by specific elements of the legislation such as abortion; while others will examine the national picture of how the political landscape is expected to change in the long run. As such, we are interested in studying the impact of a user's cognitive style on intelligence analysis tasks. Understanding a user's cognitive style is critical to determining their preferred methods for reading, remembering, learning, perceiving, and searching for information. Cognitive styles have been found to affect a user's information seeking tasks [6], the ways users interact with a graphical user interface (e.g. [5],[9]), and a user's reading task [18].

In our previous work, we discovered that there exists a consistent correlation between an analyst's actions and their final analytical conclusions from task to task [19]. In this paper, we advance this effort further by investigating how a user's cognitive style

978-0-7695-4191-4/10 $26.00 © 2010 IEEE
DOI 10.1109/WI-IAT.2010.153

may affect his/her analytical process. This involves two challenging steps: First, we need to determine a user's cognitive style in an analytical process *automatically* by analyzing the sequence of actions a user takes. Second, we need to identify the attributes of documents used/produced by the user that relate to cognitive styles.

In summary, we aim to address the following two important research questions:

1. Can a user's cognitive style be *automatically* determined based on the actions they have performed and related information while conducting an analytic task?
2. How are the actions or the information gathered in an analytic process affected by a user's cognitive style?

The first research question helps us to address the question of whether users exhibit different types of behaviors that can be explained by their cognitive style. The second question links cognitive styles and relevant documents viewed during an analytic process. This will ultimately help us address with the larger problem of: "What do we learn by knowing a user's cognitive style, what insight does this give us about analysis in general, and how does it allow us to better assist the user as they conduct an analytic task?"

Determining a user's cognitive style automatically has been a very difficult research problem. Usually, a user's cognitive style is determined by having him/her take a cognitive test such as the *Cognitive Style Analysis test* [16] or the *Study Processes Questionnaire* [4]. Even though these tests are used as the norm to determine a user's cognitive style, we encountered two major problems. First, this test is conducted separately from the process of using a target application. With the growth of offline and online information resources, it is often hard to persuade users to spend time to take a test first before using any applications. Additionally, more research effort is necessary to study whether one exhibits the same cognitive style for different cognitive processes. Therefore, there is a real need to capture a user's cognitive style dynamically as a user interacts with the target application.

There are some existing approaches that use machine learning techniques to explore navigation behaviors in a search to determine a user's cognitive styles (e.g. [7]). However, these approaches did not take into account the content associated with a user's navigation behaviors. In this paper, we propose using Hidden Markov Models (HMMs) [15] to determine a user's cognitive style based on their observed actions during an analytic process. The *novelty* of our approach is that we take the content of user knowledge (e.g. topics extracted from a user's query) into consideration while we explore a user's navigation behaviors. The content of a document is one of the most important indicators that helps determine whether or not a document is relevant to a user because relevancy is a match between a user's information needs (including queries, context, knowledge) with the content of a document ([21], [22]). Unfortunately, this factor has not been explored in depth to determine a user's searching behaviors as well as his/her cognitive style. An additional contribution of our approach is that we do not require lots of training data nor do we require users to take a separate cognitive style test as needed in other learning approaches. Our HMM approach only uses two users in the training set and it achieves an average recall of 72%. This helps to solve the cold-start problem for target applications that may use our approach in learning more about their users. Our experiment shows that the selection of training data does influence the accuracy of categorization. However, a random selection still results in an adequate recall rate on average.

Returning the two research questions above, the second question requires exploring the link between a user's cognitive style and his/her actions as well as the content of documents accessed or saved during their analytic process. We address this question with two studies: First, we study the trained HMMs of different cognitive styles and summarize the different behaviors observed from two different HMMs. Second, we conduct an experiment to find out if a user's cognitive style affects a specificity-generality factor of each retrieved snippet. A snippet contains several paragraphs from a saved, relevant document. The specificity-generality factor of a document represents the level of detail that the document contains. We choose to investigate the wholist/analytic dimension of cognitive style [16] in our study because it has been found to affect users in an information seeking task, as well as affect a user's preferences in terms of document coherency [11]. Wholists tend to process information as a whole while analytics tend to process information in parts. The results show that the analytic group focuses on snippets containing specific details much earlier than the wholist group. This is also in line with the definition of wholist and analytic groups from social science [16][8]. This experiment helps verify the classification results obtained by our HMM model.

All of the experiments in this paper use the APEX 07 collection [20], which was created by the National Institute of Standards and Technology (NIST) to simulate an analytic task in the intelligence community. This paper is organized as follows: First, we give a brief review of cognitive styles and provide readers with some background on document graphs. Next, we present our experiment on determining a user's cognitive style using Hidden Markov Models.

This is followed by an experiment to explore the relationship between a user's cognitive style and the specificity-generality factor of retrieved snippets during an analytic process. We will then conclude with the discussion of a user's cognitive styles and knowledge base and our future work.

2. Background

In this section, we review some basic literature on the history of cognitive styles that is relevant to information seeking and we provide a description of our document representation.

2.1. Cognitive Styles

We choose to discuss three different dimensions of a user's cognitive style that have been found to affect learning in both offline and online settings, as well as affect a user's preferences in using virtual environments, web browsing, and searching. They are holist/serialist [13][14], field dependence/field independence [23] and wholist/analytic [16].

The holist/serialist dimension was defined by Pask (1972) with a focus on learning style. Holists tend to use a *global* approach to learning while serialists tend to concentrate narrowly or *locally* on the details of the topics being learned [14]. This dimension can be measured using a number of different tests, for example: Free Learning technique [13] and Study Processes Questionnaire [4].

Field dependence/field independence [23] measures the degrees "to which a learner's perception or comprehension of information is affected by the surrounding environment, or fields" ([8], page 87). Field dependents may find it hard to find the information that they are looking for, given the noise and ill-defined problems that they are working on. Field independents can find ways to recognize relevant information, or make problems that they are working on more concrete. Field dependence is considered to be a global dimension while field independence focuses on the details of the fields.

Lastly, the wholist/analytic dimension [16] is closely related to field dependence/field independence dimension [23]. It reflects the preferred way that a user organizes or processes information either in its entirety (wholist) or in parts (analytic). Analytic users may have difficulty seeing the big picture when solving a problem while wholist users may have difficulty decomposing a complicated problem into smaller subcomponents. The wholist/analytic dimension can be mapped to the field dependence/field independence dimension [17]. The wholist/analytic dimension is usually measured by an appropriate computer-based test such as *Cognitive Style Analysis* (CSA) [16]. This test compares the response time of a user while he/she responds to a set of analytic or wholist questions. At the end of the test, each user will be assigned to one of these three groups: wholist, analytic, or intermediate. If this measure is a number below 1.03, it is equivalent to wholist (and also field dependent) individuals; if it is greater than 1.36, it is equivalent to analytic (and also field independent individuals). Otherwise, the individual is classified as intermediate.

These three dimensions (holist/serialist, field dependent/field independent, wholist/analytic) of a user's cognitive style essentially address the global-local issue of a user's preferences while performing information seeking, analyzing information, and problem solving. Much of the existing studies have investigated the impacts of a user's cognitive style on browsing/searching and online learning. Some examples include studies in the information seeking domain ([5], [24], [6]), online learning domain [3], and the human computer interaction domain ([3], [7]). Even though there are a lot of interests in exploring a user's cognitive style to help improve his/her performance, there is very little work on determining a user's cognitive style automatically. The existing approaches that determine a user's cognitive style require having lots of training data to start and require users to take separate cognitive style tests (e.g. [7]).

2.2. Document Representation

In our experiments presented in this paper, we represent each snippet, or query issued in an analytic process using a document graph. A snippet is a part of a document that a user has saved in analytic process. This provides evidence indicating that the user has found this piece of information relevant. A Document Graph (DG) is a directed acyclic graph (DAG) constructed based on natural language text. Within a DG, there are two types of nodes: *concept nodes* and *relation nodes*. A concept node contains a noun or a noun phrase and a relation node links two concept nodes. Two types of relations are defined – the *"isa"* relation and the *"related to"* relation. A Prepositional phrase-heuristic (PP-heuristic), a Noun Phrase heuristic (NP-heuristic), and a Sentence heuristic (S-heuristic) [12] are used to extract relationships from a sentence for inclusion in a DG. The NP-heuristic mainly defines set-subset relationships between two concept nodes and is denoted using the *"isa"* relation. This heuristic also generates "*related to*" relationship between the main noun phrase and the supporting terms in that noun phrase. Both the S-heuristic and the PP-heuristic generate "*related to*" relations. Figure 1 is an example showing how a DG would look. The DG is built from the sentence "Fissures within the clerical community do exist." Two concept nodes "Clerical Community" and "Fissure" are linked by a "related to" relation node. A "related_to" relation is identified by the NP-heuristic between "Clerical Community" and "Clerical". Concept node "Clerical Community" has an "isa" relationship with "Community".

The method we use to compare two DGs is modified from [10]. We check for a sub-graph of one DG in another DG. This method gives us similarity values between 0 and 1, with 1 meaning identical and 0 meaning totally different. The similarity is defined as:

$$Similarity(DG_1, DG_2) = \frac{n}{2N} + \frac{m}{2M}$$

where n is the number of concept nodes shared by DG_1 and DG_2, m is the number of relation nodes shared by DG_1, and DG_2. N is the total number of concept nodes in DG_1 and M is the total number of relation nodes in DG_1.

3. Our HMM Approach

In order to identify the cognitive style of a user during information seeking, we propose using Hidden Markov Models (HMMs) [15], with a user's information content and navigation information as observables. The idea is to train a model for each style using action sequences from representative users, and then to test the action sequences of other users against each model to identify their cognitive styles.

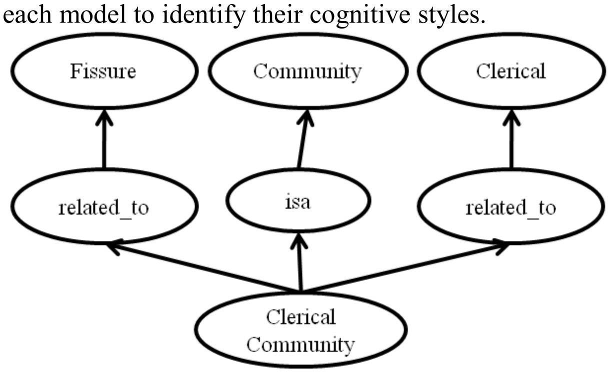

Figure 1. An example of a document graph.

3.1. Overview

An HMM is a statistical model which is used to model an *assumed* time-varying, stochastic process with unobserved states. It is described by states and output symbols. Each state has a probability distribution over the symbols. The states are hidden from the user but the output symbols of a state are visible. An HMM defines how sequences of symbols may be generated with some particular characteristics. Intuitively speaking, it can be used to predict the next action a user is going to take in a probabilistic fashion. In an HMM, the current state is determined by the last state and the emitted symbols, and the order of the symbols forms a particular HMM model. Likewise, cognitive styles can be regarded as particular ordered sequences of actions that people conduct while they switch between states as they perceive some kind of information. We use HMMs to model cognitive styles in the information searching stage of an analytic process, in which the symbols represent the actions users take. The HMMs are constructed by training representative users of each style, and then are used to categorize users in the testing set according to the likelihood of each model fitting their sequence of actions.

3.2. Dataset and Ground Truth

We use the APEX 07 dataset, which was initially used to evaluate the *IARPA Collaboration and Analyst/Effectiveness* (CASE) program's tools. Eight intelligence analysts were involved in analyzing two questions: "*whether Imar's clerical community supports the president's nuclear program"* and *"whether there is fissure between clerics*". Note that we sanitized these questions by modifying specific names and places. Each action taken by an analyst was recorded as an analysis logging event (ALE). Each ALE contains the name of the action, the content of the action (e.g. the content of the snippet being accessed), and the time of the action. The description of the APEX 07 dataset is shown in the Tables 1 and 2, respectively.

We create our ground truth by manually categorizing the analysts into wholists and analytics with reference to the characteristics in Table 3, which was created following the guidance in [16]. More specifically, we read through the ALEs of each analyst, especially the Search ALEs, study the topic of each action, and try to match their transition of topics with the patterns shown in Table 3. Four of the analysts (APEXB, APEXC, APEXH, APEXL) are categorized as wholists and the rest as analytics (APEXE, APEXF, APEXK, APEXP).

We can see from Table 3 that the order of the topics in a search is critical to identifying the style. Thus we categorize the Search ALE into three main topics which are "Imar president" (t_1), "nuclear program" (t_2) and "clerical community" (t_3) according to the content of the queries. To categorize the Search ALEs, several keywords were proposed for each topic. For instance, "nuclear program" includes the key words: cleric, fissure, divide, loyal, and so forth. We then use a simple keyword mapping technique to match each query with a key word of a topic. For queries that do not match with any keywords, we can either categorize them as *unknown search* or as belonging to the same topic as the preceding query. Since it is not always true that all analysts keep focusing on an aspect of a topic for a while before enough information is obtained or before it is found to be irrelevant, we currently encode queries with unidentifiable topics as an unknown search.

Table 1. Statistics of APEX 07 by types of actions.

ALE Type	Number of ALEs
Search	903
Access	2386
Retain	548
Assess	123
Discard	762
Make hypothesis	21
Associate evidence	207
StartApp	237
Total	5,187

Table 2. Statistics of APEX 07 by analysts

Analyst	Number of ALEs	Task starts at	Task ends at
APEXB	642	2007-12-10	2007-12-14
APEXF	482	2007-12-07	2007-12-14
APEXK	762	2007-12-07	2007-12-14
APEXC	896	2007-12-10	2007-12-14
APEXH	535	2007-12-07	2007-12-14
APEXL	614	2007-12-07	2007-12-14
APEXE	474	2007-12-07	2007-12-14
APEXP	548	2007-12-07	2007-12-14

Table 3. Comparisons of the searching behaviors of wholist and analytical styles.

	Wholist	Analytics
1	adopt a *global approach*	use a *local learning* approach
2	examine interrelationships between several topics early	overall picture emerged relatively late
3	constantly move between several topics	examine one thing at a time

3.3. HMM Model

In the HMM models, the name of each ALE action (*Access, Save, Remove, Assess, Associate Evidence*) is encoded into a numeric symbol with the exception of the Search ALE (*Search t_1, Search t_2, Search t_3, Search Unknown*) which is encoded into 4 different numerical symbols depending upon the topic of each search query. Each analyst's action sequence serves as a sample sequence to be passed to an HMM model. In the current HMM model, we consider the content of a user's query. (We plan to consider the content of the documents that analysts have accessed in our future work). In our pilot model, we use three states s_i *(i=1-3)* for simplicity. We see that after training, the states do take on apparent meanings. Because there are eight analysts in total, half of them are used as training samples and the rest are used as testing samples. In other words, two analysts are used as representatives for each cognitive style. Assume we know APEXH and APEXL are wholists and APEXE and APEXK are analytics. We then take the sample sequences of the two in each group as training samples and use the Baum-Welch algorithm [1] to train the corresponding HMM models. The impact of the choices of analysts in the training set will be described in more detail in the *categorization of searching style* section. After the HMM models are generated, we test all the analysts' sequences including the training samples against each of the models. Finally we categorize each analyst according to the log likelihood of each model.

Model description

Figure 2 demonstrates the Wholist and the Analytic HMM models generated by the training data. Figure 2(a) shows the model of a wholist and Figure 2(b) shows that of an analytic. In the wholist model, we note that when one enters a state, he/she is mostly likely to stay in that state for the next step. The table of the transfer from each state to itself is shown in the figure and the symbols with significantly high probability in each table are highlighted. $s_1 \rightarrow s_1$ happens mostly for Search actions, $s_2 \rightarrow s_2$ for Access actions, and $s_3 \rightarrow s_3$ for Save and Associate evidence actions. Therefore, we can naturally associate s_1 with the information searching stage, s_2 with the information analyzing stage and s_3 with the information selection stage.

We can identify the features for a wholist and an analytic by analyzing the corresponding HMM model. The model of a wholist can be briefly described as the following: the analyst starts by searching for a topic, and stays on the searching stage looking into various topics. Then he proceeds to the analyzing stage and accesses some documents. In the analyzing stage, he may go back to searching again or proceed to the information selection stage to save some documents and associate evidence with his arguments. After that, he finishes one cycle and starts another by going to the searching stage again to search for more detailed information. In summary, the features from the wholist model are: (i) they switch between topics; (ii) they select documents early; and (iii) they associate evidence frequently. In contrast, in s_1 of the analytic model, they both search t_1, t_3 and access documents. In s_2 they search t_2 and access documents; while they save documents in the last stage. It is obvious that analytics tend to separate the searching and analyzing stages of each topic. In addition, the probability of transferring from the analyzing stage to the information selection stage is small, which means analytics spend more time on the analyzing stage. Most of the features of the wholist and the analytic styles are explained quantitatively in the model. Specifically, it describes how much focus the analyst puts on a topic, how the focus is transferred, how the analyst proceeds to different stages of analysis, and how long a cycle from searching to analyzing is.

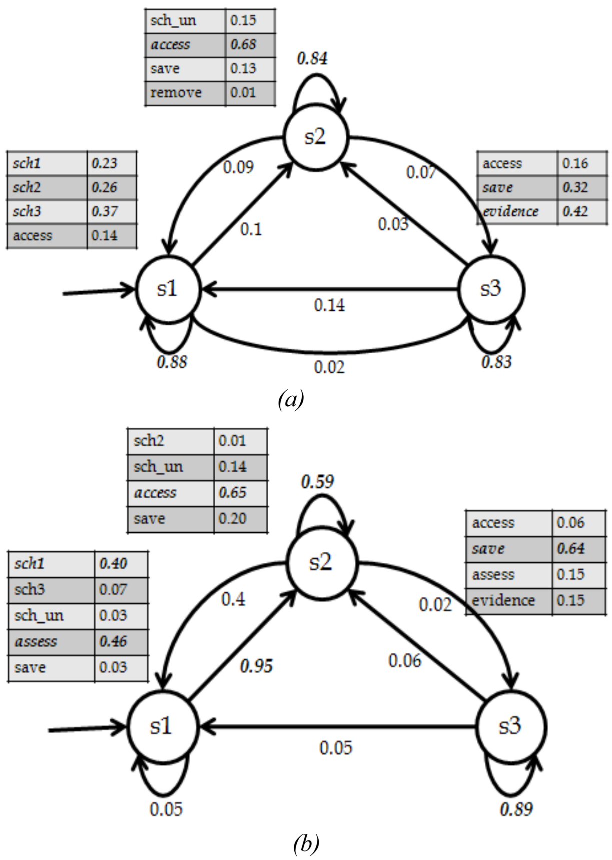

Figure 2. HMM models trained with the Baum-Welch algorithm. (a) Wholist HMM model. (b) Analytics HMM model.

Categorization of searching style

Table 4 lists the log likelihood of each analyst's sequence against the HMM models, which represents how well the sequence fits a model. Since each analyst has more than 300 symbols in his sequence, and the likelihood of each sequence is the joint probability of all symbols given the model, the values of the log likelihoods look negatively high. The values are plotted in Figure 3. Black dots represent analysts manually classified as wholists, and grey dots represent analysts manually classified as analytics. The figure shows that four out of four analytics are correctly classified and three out of four wholists are correctly classified. Thus, the recall rate for wholists is 75%, the recall rate for analytics is 100%, and the overall recall rate is 87.5%.

Table 4. Log likelihood of each analyst's sequence against the wholist model (W) and the analytic model (A)

	B	C	E	F	H	L	K	P
W	-445	-558	-429	-569	-477	-618	-550	-461
A	-471	-468	-372	-453	-560	-480	-767	-454

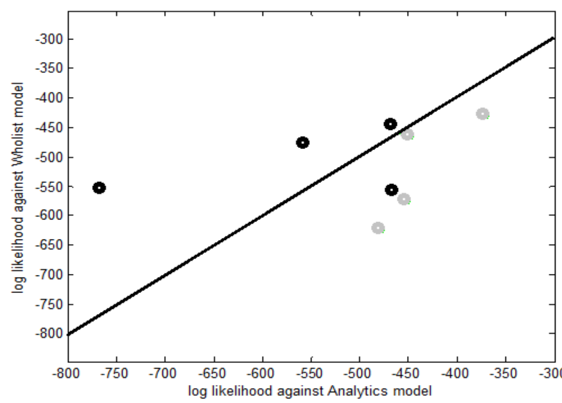

Figure 3. Log likelihood in analytic model against the log likelihood in wholist model.

Since it is not clear whether the high recall rate is due to lucky selection of the training sample, we conducted an exhaustive training process. In particular, we selected any two wholist combined with any two analytic as our training sample, which form 36 cases altogether. Then for each case of the training sample, we build models of wholist and analytics based on them, classify all the analysts' sequences by the models, and calculate the overall recall rate. During the experiment, we observed that the log likelihoods of some test samples against a model can be negatively infinite. It was because the training samples used to train the model exhibit some unique action sequences that cannot be found in the test samples. If the log likelihoods of a test sample against both models are negatively infinite, we will not count it as a successful classification. The result of the test is plotted in Figure 4. The mean recall rate is 0.72, with the highest value 0.875 and lowest value 0.375. The lowest recall rate was found in one case, in which 4 out of 8 analysts have the log likelihoods against both models being negatively infinite. It means at least one of the training samples for each model, which are APEXB, APEXH for wholist and APEXF, APEXP for analytics, exhibit unique patterns in their ALEs. If we had used only one training sample for each model, it would be more likely to produce unique patterns. Although the lowest recall rate is somewhat low, the majority of the cases have a recall rate around 0.75. Intuitively speaking, although the selection of training samples does influence the classification result, a random selection on average would promise a relatively high recall rate (0.72) indicating the high recall rate achieved is not simply due to lucky selection of the training sample.

4. Cognitive Style and Semantics of a document

In the previous sections, we built an HMM model to determine a user's cognitive style. Now, we would like to verify if a user's cognitive style, as determined above, agrees with their definitions in [16] by exploring the link between a user's cognitive style and attributes of the content of a document. This specific/general attribute of documents will help guide us in the process of automatically determining a user's cognitive styles.

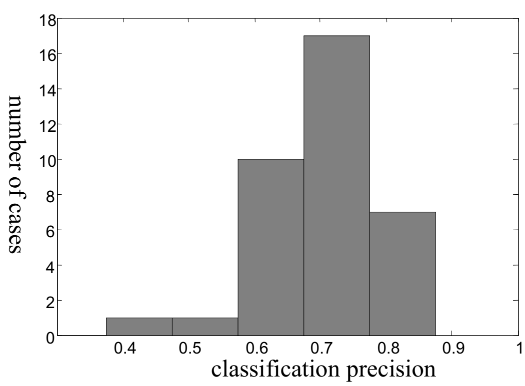

Figure 4. Number of cases with a specific recall rate of classification

We have chosen one aspect of document content that is potentially linked to a user's cognitive style – the *specificity-generality* factor. The specificity-generality factor of a document represents the level of detail of the information that particular document contains. The basis for our choice of this factor is that a user's cognitive style, such as wholist/analytic or field dependent/field independent, essentially addresses the global-local issue of a user's preferences while solving a task at hand. This factor directly relates to the global-local issues in that it can be used to determine whether a user focuses locally on specific details or globally on conceptual information. Therefore, it is intuitive to explore whether there exists a significant difference in specificity factors between wholists and analytics.

In this set of experiments, we explore if there is any difference in terms of specificity-generality factors of any snippets collected by wholist and analytic analysts. We use two groups of analysts classified manually and by using HMM models. We have one dependent variable which is the specificity-generality factor of a document. We compute this factor as follows: (i) we create a common knowledge base for each analyst that contains *"isa"* relationships and *"related_to"* relationships generated from NP-heuristics from all document graphs representing the saved snippets. This common knowledge base for each analyst contains his own domain knowledge represented as the concepts and relationships between concepts. We choose *"isa"* relationship because it reflects clearly the set (general)-subset(specific) relationship. The "*related_to*" relations generated from NP-heuristics also represents the relations between the main noun phrases (more specific) to individual supporting terms (more general). (ii) We compute this factor for each snippet saved by each analyst as follows: $factor(d_i) = \frac{\sum_c l(c)}{N}$ in which c is any concept nodes found in this document d_i and the corresponding common knowledge base of that analyst. The node c also needs to have either parents or children. N is total number of all such a node c in the document d_i. The level of a node $l(c)$ is computed as follows:

$$l(c) = \begin{cases} 0 \; if \; c \; has \; no \; parents \\ \max\left(l\big(p(c)\big)\right) + 1 \;\; otherwise \end{cases}$$

in which $p(c\,)$ is a parent of the node c. If this factor is small, the document contains more specific and detailed information while if this factor is big, the document contains more general information.

We perform this experiment with the wholist/analytic groups that were determined manually and with those determined by the HMM models as mentioned above. For each analyst, we use all possible saved snippets. The average specificity-generality factor of the analytic group is always smaller than the wholist group. We use one-way ANOVA to find out if the mean difference is statistically significant. For the wholist/analytic groups determined by manual classification in which APEXB, APEXC, APEXH, APEXL are classified as wholists and APEXE, APEXF, APEXK, APEXP are classified as analytics, the average specific-generality factor for the analytic group was 3.0511 while the average for the wholist group was 3.389. We found that the difference of these averages was statistically significant (n=614, P-value < 0.05), as shown in the Table 5. In other words, a user's cognitive style affects the type of documents that a user will explore next. Analytic analysts focus on documents with significantly more specific and detailed information than wholist analysts. For the wholist/analytic groups determined by HMM classification in which APEXB, APEXH, APEXL were correctly classified as wholists and APEXE, APEXF, APEXK, and APEXP were correctly classified as analytics, the average for the analytic group was 3.051 while the average for the the wholist group was 3.251. We found that the difference of these averages between two groups is also statistically significant (n=437, P-value = 0.039 < 0.05), as shown in the Table 6. This experiment shows that there is a relationship between a user's cognitive styles with specificity-generality factor of the contents of the documents that are retained in an analytic process. This can be used as a guiding factor in combination with the classification results from our HMM model. It can also be used to verify the classification of a group of users' cognitive styles.

Table 5. ANOVA analysis for groups determined by manual classification

	Sum of Squares	df	Mean Square	F	Sig.
Between groups	17.448	1	17.448	22.792	.000
Within groups	468.497	612	.766		
Total	485.945	613			

Table 6. ANOVA analysis for groups determined correctly by HMM

	Sum of Squares	df	Mean Square	F	Sig.
Between groups	3.213	1	3.213	4.297	.039
Within groups	325.311	435	.748		
Total	328.524	436			

5. Cognitive Styles and User's Knowledge Base

As Section 4 studies the impacts of one's cognitive styles on the choice of a document, this section studies the impacts on the preferences of topics to explore. Does knowing one's cognitive style helps us better understand why and how an analyst explores topics? Here, we propose to analyze the growth of the analysts' knowledge base with respect to different styles.

In order to capture the knowledge base of analysts, we convert all the saved documents into document graphs and aggregate these graphs over time to form a semantic representation of the information they have perceived and learned as time goes by, and to analyze how one's knowledge base evolves over time with respect to one's cognitive style. We assume that the wider a document graph is, more topics are covered, and that the higher a document graph is, more detailed the document is. Therefore, for wholists who access multiple topics at once, the document graph tends to be wide and short (high specificity-generality factor); while for analytics, the document graph is high and thin (low specificity-generality factor).

Based on the above assumption, we studied the content of the documents in the APEX 07 data. Since the analysts are more certain about the relevancy of the documents that they saved rather than the documents they accessed, we converted only the saved documents into document graphs and measured their widths and heights. Our hypothesis is that by comparing the growth of width and that of height over time, we should see width grow fast at first with height catching up later for wholists, and vice versa for analytics. We plotted the widths and heights over time for analysts APEXB, APEXC, APEXF and APEXK in Figure 5. Although the plots do not follow the hypothesis exactly, they exhibit a similar pattern, in which for both styles width grows faster at first but the wholists start with relatively wide DGs and the width of the DGs seems to grow slower than the analytics. Therefore, APEXB and APEXC follow the wholist pattern and APEXF and APEXK follow the analytic pattern, which is consistent with our manual classification. The results also match observations from the trained HMMs in the earlier section. In conclusion, an analytical user expands his knowledge bases by exploring more detail from familiarized topics. On the other hand, a wholist users expands his knowledge base by constantly exploring new topics.

6. Conclusion and Future work

We presented our current efforts on identifying cognitive styles using HMMs to represent and classify searching styles of analysts and using specificity-generality factor as well as width and height factor of a document to verify this classification. Our HMM model can be used to determine a user's cognitive styles while using Web applications automatically.

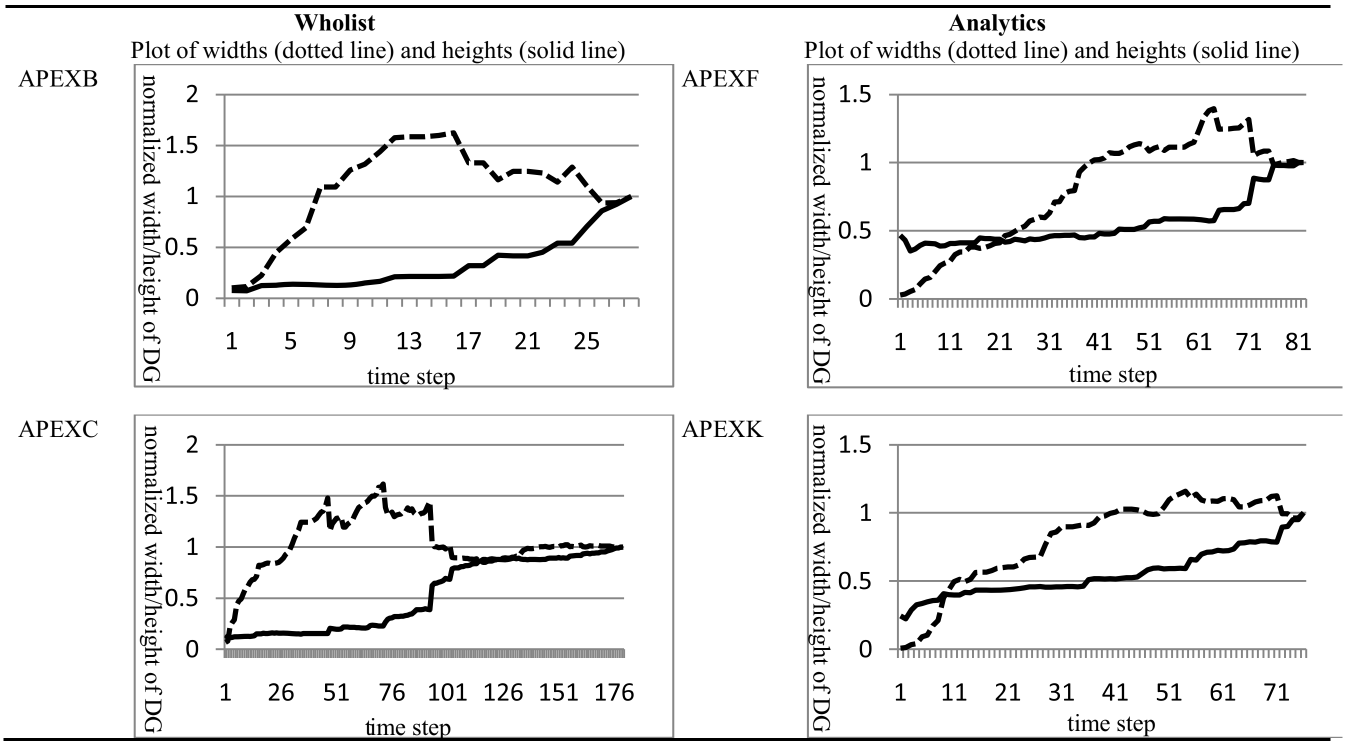

Figure 5. Plot of width and height of the saved DGs against time step for analysts APEXB, APEXC, APEXF and APEXK

Searching keywords, visited pages, and implicit relevant feedback can be used as inputs for such a model. Additionally, the content and presentation of the summaries of searching results can be personalized according to a user's cognitive styles to improve user satisfaction. We also studied how cognitive styles influence the growth of the knowledge base of the analyst. We found out that for wholists, they start with a relatively wide DG, which covers more topics, and quickly dive into much wider DGs; while for analytics, they start with a tall DG, which includes details of a single topic, and the coverage of topics grows as they dive into more details.

The results seem promising, but there is still room for improvement. First, a more sophisticated natural language processing tool, such as Latent Dirichlet Allocation [2], should be used in classifying the topics. The current key word mapping method does not guarantee that a query has the same semantic meaning as a topic, and some queries' topics cannot be recognized successfully, which results in the loss of information about state transfer due to the unknown topics. Second, another way to build the wholist and analytic models is to manually design the models based on the characteristics both observed from the models we constructed in the pilot study and from psychologists. The advantage is that we do not need any training sample so that all the analysts can be used as testing samples. The difficulty exists in how to validate the model. An alternative would be to build a model for each analyst and assume the two most different ones to represent the wholist and analytic models, and use them to classify the rest of the analysts. However, a potential problem is that the result varies when an analyst switches group. Another observation from our result is that the test not only classifies all the analysts but also measures the likelihood of each style. Therefore, a possible byproduct of our method is a score to indicate the level of wholist/analytic tendencies in the style of each analyst. Third, specificity-generality and height-width factors can be implemented and used in combination with our HMM model to determine a user's cognitive styles automatically. Fourth, more information about the document/snippet content can be included into the HMM models. In the current work, we explicitly encode search queries on different topics into different symbols. It is also useful to include other attributes of a document such as the length of the saved document (e.g. wholists tend to access long documents and the opposite for analytics), and the retrieval of the specificity of the accessed document (e.g. wholists tend to access documents with general terms while analytics tend to access documents with specific terms). Last, a comprehensive evaluation should be conducted to justify the effectiveness of our HMM model on a larger testbed.

Acknowledgments. This work was supported in part by Air Force Office of Scientific Research, Grant No. FA9550-07-1-0050 and by a grant from IARPA. Special thanks to Brittany Clark for comments on improving this paper.

References

1. Baum L. E., Petrie T. , Soules G. , and Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41 (1), pp. 164-171. 1970.
2. Blei, D. M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, J. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3,pp. 993–1022. 2003.
3. Chen, Y. S. and Macredie, D. R. Cognitive styles and hypermedia navigation: Development of a learning model. *Journal of the American Society for Information Science and Technology,* 53(1), pp. 3-15. 2002.
4. Ford, N. Learning styles and strategies of postgraduate students. *British Journal of Educational Technology,* 16, pp. 65– 77. 1985.
5. Ford, N., and Miller, D. Gender differences in Internet perception and use. In *Proceedings of the Third ELVIRA Conference on Electronic Library and Visual Information Research (ASLIB), London.* 1996.
6. Ford, N.; Wilson, T. D.; Foster, A.; and Ellis, D.(2002). Information Seeking and Mediated Searching. Part 4. Cognitive Styles in Information Seeking. *Journal of the American Society for Information Science and Technology,* 53(9), pp. 728-735. 2002.
7. Frias-Martinez, E.; Chen, Y. S.; Liu, X. Automatic cognitive style identification of digital library users for personalization. *Journal of the American Society for Information Science and Technology,* 58 (2), pp. 237-251. 2007.
8. Jonassen, D. and Grabowski, L. B. *Handbook of Individual Differences, Learning, and Instruction*. Lawrence Erlbaum. 1993.
9. Lee, J. and Boling, E.. Information-conveying approaches and cognitive styles of mental modeling in a hypermedia-based learning environment. *Journal of the American Society for Information Science and Technology,* 59(4), pp. 644-661. 2008.
10. Montes-y-Gómez, M.; Gelbukh, A.; and Lópes-López, A. Comparison of Conceptual Graphs. In *Proceeding of MICAI-2000, 1st Mexican International Conference on Artificial Intelligence.* 2000.
11. Nguyen, H.; Santos, E. Jr.; Russell, J.; and Smith, N. Evaluation of the Effects of User-Sensitivity on Text Summarization. In *Proceedings of 2008 IEEE/WIC/ACM International Conference on Web Intelligence.* Sydney, Australia, pp. 927-931. 2008.
12. Nguyen, H. Capturing User Intent for Information. Ph.D. Dissertation, University of Connecticut. 2005.
13. Pask. G.; Scott, B, C, E. Learning strategies and individual competence. *International Journal of Man-Machine studies,* 4(1), pp. 217-253. 1972.
14. Pask. G. Styles and Strategies of Learning. *British Journal of Educational Psychology.* 46(II), pp. 128-148. 1976.
15. Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Readings in speech recognition,* pp. 267 - 296. 1990.
16. Riding, R. J.; and Cheema, I. (1991). Cognitive styles - An overview and integration. *Educational Psychology*. 11 (3/4), pp. 193-215. 1991.
17. Riding, R. J.; and Rayner; S. G. *International Perspectives on Individual Differences, Volume 1: Cognitive Styles*. Ablex Publishing. 2000.
18. Rosa, M. H. Relationships between Cognitive Styles and Reading Comprehension of Expository Text of African American Male Students. *The Journal of Negro Education Pedagogical and Contextual Issues Affecting African American Males in School and Society.* 63(4), pp. 546-555. 1994.
19. Santos, E. Jr.; Nguyen, H.; Yu, F.; Kim, K.; Li, D.; Wilkinson, J.; Olson, A.; Russell, J. Intent-driven insider threat detection in intelligence analyses. In *Proceedings of 2008 IEEE/WIC/ACM International Conference on Intelligent Agent Technology,* Sydney, Australia, pp. 345-349. 2008.
20. Santos, E. Jr.; Nguyen, H.; Yu, F.; Kim, K.; Li, D.; Wilkinson, J.; Olson, A.; Russell, J. Capturing User Intent for Analytic Process. In *Proceedings of User Modeling, Adaptation, and Personalization (UMAP 09)*, pp. 349-354, Trento, Italy. 2009.
21. Saracevic, T. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology,* 58(3), pp. 1915-1933. 2007.
22. Saracevic, T. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology,* 58(13), pp. 2126-2144. 2007.
23. Witkin, H. A.; Moore, C. A.; Goodenough, D. R. , and Cox, P. W. Field dependent and field independent cognitive styles and their educational implications. *Review of Educational Research*, 47(1), pp. 1-64, 1977.
24. Wood, F.; Ford, N.; Miller, D.; Sobczyk, G.; and Duffin, R. Information skills, searching behaviour and cognitive styles for student centred learning: A computer-assisted learning approach. *Journal of Information Science*, 22(2), pp. 79-92. 1996.