Automatic Detection of Search Tactics in Collaborative Exploratory Web Search Process

Zhen Yue
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15213
zhy18@pitt.edu

Shuguang Han
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15213
shh69@pitt.edu

Daqing He
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, PA 15213
dah44@pitt.edu

ABSTRACT
Information seeking process is an important research topic in information seeking behavior. Collaborative information seeking (CIS) has attracted many researchers' attention in recent years, but the investigation of CIS process is still rare. Investigations on search processes can either be macro-level or micro-level. The macro-level investigation focuses on establishing theoretical models while micro-level investigation focuses on identifying descriptive categories such as user action or search tactics. In this paper, we proposed an automatic technique and explicitly model the latent search tactics using a Hidden Markov Model. HMM results show that the identified search tactics transition patterns in individual information seeking process are consistent with Marchionini's information seeking process model. Then, we applied the HMM in CIS and found different patterns of search tactics compared to the individual search. With the advantages of showing the connections between search tactics and search actions, and the transitions among search tactics, we argue that Hidden Markov Model is a useful tool to investigate information seeking process, or at least it provides a feasible method to analyze large scale dataset.

Categories and Subject Descriptors
H.3 INFORMATION STORAGE AND RETRIVAL H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – Collaborative computing, Computer-supported cooperative work

General Terms
Experimentation; Human Factors

Keywords
Collaborative information behavior; Exploratory Search; Hidden Markov Model; Information Seeking Process

1. MOTIVATIONS AND BACKGROUND
Information seeking process is one of the major topics in information seeking behavior research. In individual search, researchers had employed two major approaches to investigate information seeking process. One is modeling macro-level information seeking process, which focuses on qualitative constructs such as stages and context in information seeking process. Kuhlthau's ISP model [6] and Marchionini's [8] ISP model both took such kind of approach. The other one is modeling micro-level information seeking process by identifying descriptive categories such as user action, search strategies or search tactics and the transition relationships among them [5]. One study that took this approach is [14], in which the researchers investigated the transition patterns of search tactics at different phases within one search session.

Collaborative information behavior is a relatively new research area compared to individual information behavior research. Investigating collaborative search process is crucial for designing and evaluating systems that support collaborative information seeking. Shah and Gonzalez-lbanez [12] attempted to map Kuhlthau's ISP model to collaborative information seeking. To the best of our knowledge, there is no work had focused at micro-level collaborative information seeking process by identifying collaborative search tactics and the transition relationships among them.

Search tactic has been recognized as a mean of examining search processes. Bates [2] proposed the notion of search tactics which consist of a move or moves applied to advance the search process. She proposed a model including 32 search tactics in four categories. There are also many other framework of search tactics had been proposed. Xie and Joo [14] manually coded user search transaction logs according to a predefined framework including 13 search tactics. We can see most of previous researches highly rely on predefined framework of search tactics and manually coding, which makes it difficult to be expanded or used in a different or large-scale dataset. However, there is no existing widely-recognized collaborative search tactics model or framework in collaborative information seeking. The search tactics defined in individual information seeking cannot be simply applied in collaborative environment because user actions involved in the process of collaborative exploratory search are more complicated than that in individual search. In collaborative search, users do not only need to take actions toward the completion of search task, but also need to take actions to facilitate the collaboration. Therefore, revealing the search tactics behind user actions is a challenging task.

Automatic methods have been explored in some work. Chen and Cooper [3][4] used both stochastic model and clustering techniques to examine search tactics in a Web-based library catalog. However, they usually missed explain the latent rationale behind the search tactics. Their identified search tactics are simply the aggregation of sequential behaviors while the connections among user actions and search tactics are missing. In this paper, by treating the sequence of user actions as Markov chains, we modeled users' search tactics explicitly as hidden variables. In this way, we propose using Hidden Markov Model (HMM) to automatically uncover the relationship between users' actions and search tactics. The HMM algorithm is used to identify the hidden search tactics, their connections with user actions is output in the emission probabilities. The relationships among search tactics can also be output in the transition probabilities.

2. EXPERIMENT DESIGN
Our study was designed as a set of control experiments with human participants using CollabSearch[1][15], a collaborative search system developed by the authors.

[1] http:// crystal.sis.pitt.edu:8080/CollaborativeSearch/

2.1 Experiment Conditions

We included both individual search and collaborative search in our experiments. There are two reasons for us to involve individual search: 1) the individual search results can be served as a baseline for the comparison with collaborative web search; 2) the individual web search results is used to validate our proposed HMM model. The individual information seeking behavior is well-studied and there have been several existed models. In our study, we also used one of the well-known models to validate our propose model.

As a result, our experiment has two different conditions - the Collaborative Web Search condition (COL) and the Individual Web Search condition (IND) described as follows:

COL: In this condition, two participants form a team and they worked on the same task simultaneously. As we were trying to simulate remotely-located collaboration, the participants in the same team could communicate with each other by sending instant text messages or reading each other's search histories and the collected results shared in team workspace, but no face-to-face communication was allowed.

IND: Individual search. In this condition, we had a participant work on the exploratory search tasks individually.

2.2 Participants

24 participants were recruited from the University of Pittsburgh for this study. Among them, 10 are female and 14 are male. All the participants are students and they use computers on a daily basis. 13 participants are graduate students whereas the other 11 are undergraduates. According to a question asking them to rate their search experiences from 1-7 with 1 as the least experienced and 7 as the most experienced, the response range from 4-7, thus most of our participants are experienced searchers. 16 of the 21 participants worked under the COL condition. These 16 participants signed up as pairs, and the members of each pair know each other before the study so that it was reasonable easy for them to form a team. Therefore we have 8 pairs of participants worked as 8 teams in COL. The rest 8 participants were assigned to individual search condition.

2.3 Search Tasks

Two exploratory web search tasks were used in this study. Both of them had been used in other collaborative web search studies [11] [13], so their validity for collaborative search has been examined before. One task (T1) is related to academic work, which asks participants to collect information for a report on the effect of social networking service and software [13]. The other task (T2), which is about leisure activities, asks participants to collect information for planning a trip to Helsinki [11]. Morris' [10] identified that travel planning and academic literature search are two common collaborative search tasks. Therefore, both tasks here are representative in studying collaborative web search. The task description carefully states the kind of information that the participants need to collect and the goal is to collect as many relevant snippets as possible.

2.4 Experiment Procedure

The experiment procedure was: experiments for COL condition were conducted first. Each team in COL worked on both tasks. The order of the two tasks was rotated to avoid the learning and fatigue effect. During the experiment, after being introduced to the study and the system, and filling out an entry questionnaire to establish their search background, these participants worked on a training task to get familiar with the system for 10 minutes. Then they worked on task 1 or task 2, depending on the task order assigned for each team. They had 30 minutes for each task. At the end of each task, each of them also worked on a post-search questionnaire collecting information about their satisfaction with the search results. Before the end of the experiment, participants were asked several open-ended questions for their experience with both tasks. The IND experiments were conducted after the COL experiments. The experiment procedure in IND is identical to the COL condition.

2.5 Categorizing user search actions

In order to analyze the transaction logs, we categorized user actions into 6 categories: Query, View, Save, Workspace, Topic and Chat, whose details are listed in Table 1.

Table 1: User search actions

Actions	Descriptions
Query (Q)	A user issues a query or clicks on a query from search history.
View (V)	A user clicks on a result in the returned result list
Save (S)	A user saves a snippet or bookmarks a webpage
Workspace (W)	A user clicks or edits or comments an item saved in the workspace
Topic (T)	A user clicks on the topic statement or leaves comments
Chat (C)	A user sends an message or views the chat history

3. HMM METHOD

3.1 Modeling Search Tactics

In this study, we introduced a hidden Markov Model (HMM) method to model search tactics and search actions simultaneously. The model is described in Figure 1. We have a sequence of user actions from A_1 to A_M, and each action is one of those predefined six actions: {Q, V, S, W, T, C}. Using HMM, we need to assume that we also have a sequence of hidden search tactics, from s_1 to s_M. HMM assumes that each action is generated by a corresponding hidden search tactic, but different actions can be generated by the same search tactic with different probabilities. In this case, each action is corresponding to only one search tactic, and the search tactic sequence forms a Markov Chain.

A HMM model has several parameters: the number of hidden states N (search tactics in this case), the start probability of each states $\boldsymbol{\pi}$, the transition probabilities among any two hidden states $\boldsymbol{A_{ij}}$ and the emission probability from each state to each action $\boldsymbol{b_{ij}}$. By only defining the N and $\boldsymbol{\pi}$, a Baum-Welch algorithm [7] can be used to estimate the emission and transition probabilities.

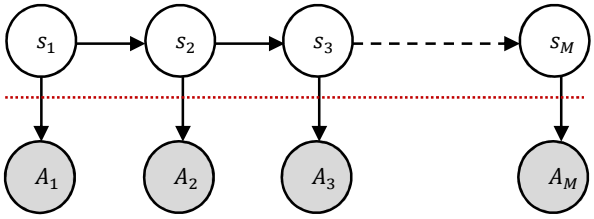

Figure 1: A Hidden Markov Model for Search Tactics

3.2 Parameter Selection

It is still an open issue for determining the number of hidden states. Determining number of hidden states **N** is a model selection problem in learning the Hidden Markov Model. A complex model with large number of states will help to increase the sequence likelihood because there are more parameters that can be used to describe the model more precisely. But it has high

risk to cause over-fitting. A simple model is less likely to over-fit on the given dataset, but it may not be able to uncover the natural feature of datasets. In model selection, the information criterion such as the Akaike information criterion (AIC) or its variants [1] and Bayesian information criterion (BIC) [9] can be used to determining the optimal number of states. In this paper, we used BIC because it also considers the sample size.

Suppose the number of parameters in HMM is p, and the number of total samples are s. The BIC is defined in Formula (1), in which L denotes the log-likelihood of all samples. p can be calculated using $p = (N-1) + (N-1) \times (N-1) + N \times (M-1)$, considering the summation of all probabilities is 1. The M denotes the number of action types. A large log-likelihood and less parameter are preferred for BIC.

$$BIC = -2 \times log(L) + \log(s) \times p \quad \text{Eq. (1)}$$

4. RESULTS

4.1 Results of IND

Model selection is the first step of analyzing HMM result. Figure 2 plots the BIC values against the number of hidden states in IND condition. We can see that BIC has the optimal value when the number of states is set to 5.

There are two different types of output from HMM: the emission probability of hidden stats and the transition probability of hidden states. Therefore, each hidden state can be represented by the emission probability distribution over user actions. The results are shown in Table 2, in which we removed the probabilities that are smaller than 0.05 for better visualizing each search tactic. S1 has a very high probability of generating the Query action. The probability of generating Save action is 0.97 for S3. S5 has 0.64 probability of generating Workspace action and 0.32 probability of generating Topic action. It may represent a search tactic for defining search problem. Although S2 and S3 seem to be the same tactic because they both have a high probability of generating the View action, they represent different search tactics due to the difference in transition probability.

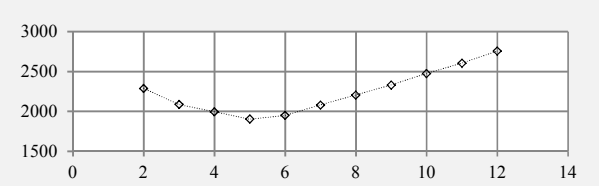

Figure 2: BIC Evaluation of HMM parameters in IND

Table 2: Search Tactics and Emission Probability in IND

	Q	V	S	W	T
S1	0.92				0.06
S2		0.97			
S3		0.98			
S4			0.97		
S5				0.67	0.32

Transition probabilities among different hidden states are another type of important output from HMM, which is shown in Figure 3. Each cell in the visualization denotes the transition probability from the row search tactic to the column search tactic. The darker the cell is, the larger the transition probability. We can S2 and S3 see have very different transition patterns. S3 has a high probability of transmitting to S4 (saving results) while S2 has a high probability of transmitting to S3. Therefore we think that S3 -> S4 represents examine a search result and then save it, S2 -> S3 represents examine a list of search result without saving.

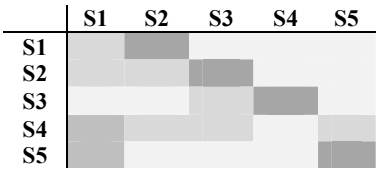

Figure 3: Transition of Search Tactics in IND

Table 3: Mapping from sub-process to HMM patterns

Sub-processes	Patterns
Define Problem	P5
Select Source	P1
Formulate Query	
Execute Query	
Examine Results	P2, P3
Extract Information	P4
Reflect/Iterate/Stop	P5

We also find that the transitions shown in Figure 3 are very similar to the transitions defined in Marchionini's model. The default transition from Marchinoinini's model can be converted into S5 → S1 → (S2 → S3) → S4, which are almost the darkest areas shown in Figure 3. Similarly, ISP model also described the high and low probability among different sub-processes. For example, "*extract information"* (S4) has high probability of transit to "*examining results*" (S2, S3) and "*formulate query*" (S1). Through the comparison, we established certain validity of HMM method in analyzing information seeking process. The mapping from HMM result to Marchinoinini's ISP model is shown in Table 3.

4.2 Results of COL

From figure 4, we can see the BIC has the optimal value when the number of hidden state is set to 6 in COL condition.

The emission probability and transition probability are shown in Table 4 and Figure 5. Several of the identified tactics are the same as in IND condition, such as S2, S3 and S4 in COL are very similar to that in IND. However, the rest search tactics are different. Not surprisingly, we identify a new search tactic S6 has a high probability of generating Chat action. However, the influence of Chat action is not only existing in S6, but also embedded to other search tactics. For example, the identified pattern S1 in IND is mainly about issuing a query. S1 in COL is now embedded with chatting behavior. It indicates that the explicit communication between participants do influence their query behavior. Same situation exist in S5, which has a 0.17 probability of generating Chat action. This is evidence showing that the communication also influences the problem definition in the search process. It is easy to understand that participants may discuss what information to search.

In terms of transition probabilities between the hidden tactics, there are also similarities and differences between the COL and IND conditions. The similarity is that in both COL and IND, the transitions of *S*1 -> *S*2, S2 ->S3 and S3 -> S4 and are all very high. This indicates a typical pattern of Web search behavior - the participant first issues a query, then views the returned results, collects the result if it's relevant or continues viewing other results if it's not relevant. S4 in IND has a high probability of transmit to S1 while S4 in COL has a high probability of transmit to S5. This may indicate that after saving a result, participant in IND tends to issuing another query while participant in COL might discuss the saved results or what else to search with their partner.

We further identified that S1, S2, S3 and S4, both in IND and COL, are task-oriented search tactics because they are essential work of completing the Web search task. S5 in IND, S5 and S6 in COL, are support-oriented search tactics. Although they are not directly related to search, they provide indispensable support for the search. The COL condition obviously has more support-oriented search tactics compared to IND.

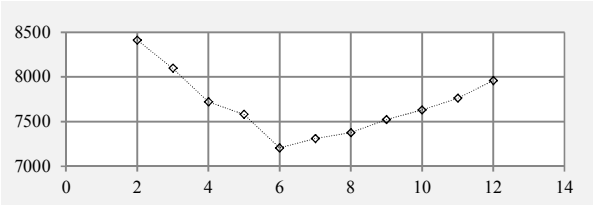

Figure 4: BIC Evaluation HMM parameters in COL

Table 4: Search tactics and transition probability in COL

	Q	V	S	W	T	C
S1	0.88					0.11
S2		0.97				
S3		1.00				
S4			0.92		0.05	
S5				0.58	0.25	0.17
S6					0.08	0.88

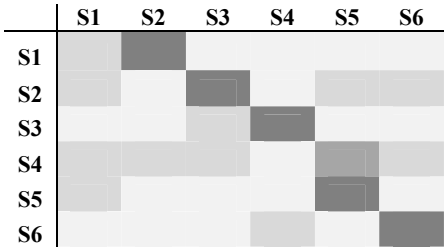

Figure 5: Transition of search tactics in COL

5. CONCLUSION

In this paper, we propose a HMM method for automatically detect search tactics in the information seeking process. A user study is conducted to compare the search tactics in collaborative exploratory search process and individual exploratory search process. We found different patterns of search tactics under collaborative and individual search conditions. The transition pattern of search tactics in individual condition is very similar to Marchoinini's information seeking process model, which to some extent validate the HMM method. Further studies are needed to validate this method as a way to analyzing collaborative information seeking process. More importantly, how to interpret the output of HMM in a meaningful way is an issue we need to address in the future study.

6. REFERENCES

[1] Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 19, 6 (1974), 716–723.

[2] Bates, M.J. 1979. Information search tactics. *Journal of the American Society for Information Science*.

[3] Chen, H.-M. and Cooper, M.D. 2002. Stochastic modeling of usage patterns in a web-based information system. *Journal of the American Society for Information Science and Technology*. 53, 7 (2002), 536–548.

[4] Chen, H.-M. and Cooper, M.D. 2001. Using clustering techniques to detect usage patterns in a Web-based information system. *Journal of the American Society for Information Science and Technology*. 52, 11 (2001), 888–904.

[5] Kim, J. 2009. Describing and Predicting Information-Seeking Behavior on the Web. *Journal of the American Society for Information Science and Technology*. 60, 4 (2009), 679–693.

[6] Kuhlthau, C.C. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*. 42, 5 (Jun. 1991), 361–371.

[7] L.E.Baum et al. 1970. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist*. 41, 1 (1970).

[8] Marchionini, G. 1995. *Information seeking in electronic environments*. Cambridge University Press.

[9] McQuarrie, A. and Tsai, C. 1998. *Regression and Time Series Model Selection*.

[10] Morris, M. 2008. A survey of collaborative web search practices. ACM.

[11] Paul, S.A. and Rosson, M.B. 2010. UNDERSTANDING TOGETHER : SENSEMAKING IN COLLABORATIVE INFORMATION SEEKING by. May (2010).

[12] Shah, C. and González-ibáñez, R. 2010. Exploring Information Seeking Processes in Collaborative Search Tasks Chirag Shah. *ASIS&T* (2010).

[13] Shah, C. and Marchionini, G. 2010. Awareness in Collaborative Information Seeking. *Journal of the American Society for Information Science and Technology*. 61, 10 (2010), 1970–1986.

[14] Xie, I. and Joo, S. 2010. Transitions in Search Tactics During the Web-Based Search Process. *Journal of the American Society for Information Science*. 61, 11 (2010), 2188–2205.

[15] Yue, Z. et al. 2012. A Comparison of Action Transitions in Individual and Collaborative Exploratory Web Search. *The eighth asia information retrieval societies conference (AIRS2012)* (2012).