Learning Dynamic Information Needs: A Collaborative Topic Variation Inspection Approach

I-Chin Wu
Department of Information Management, Fu Jen Catholic University, Taipei, 242 Taiwan.
E-mail: icwu.fju@gmail.com

Duen-Ren Liu and Pei-Cheng Chang
Institute of Information Management, National Chiao Tung University, Hsinchu, 300 Taiwan.
E-mail: dliu@iim.nctu.edu.tw; mrkid.tw@gmail.com

For projects in knowledge-intensive domains, it is crucially important that knowledge management systems are able to track and infer workers' up-to-date information needs so that task-relevant information can be delivered in a timely manner. To put a worker's dynamic information needs into perspective, we propose a topic variation inspection model to facilitate the application of an implicit relevance feedback (IRF) algorithm and collaborative filtering in user modeling. The model analyzes variations in a worker's task-needs for a topic (i.e., personal topic needs) over time, monitors changes in the topics of collaborative actors, and then adjusts the worker's profile accordingly. We conducted a number of experiments to evaluate the efficacy of the model in terms of precision, recall, and F-measure. The results suggest that the proposed collaborative topic variation inspection approach can substantially improve the performance of a basic profiling method adapted from the classical RF algorithm. It can also improve the accuracy of other methods when a worker's information needs are vague or evolving, i.e., when there is a high degree of variation in the worker's topic-needs. Our findings have implications for the design of an effective collaborative information filtering and retrieval model, which is crucial for reusing an organization's knowledge assets effectively.

Introduction

Information seeking or searching is regarded as the primary activity of knowledge workers when they execute tasks. The 2004 International Data Corporation (IDC) Report (Feldman, 2004) estimated that 90% of a company's accessible information is only used once. If knowledge cannot be accessed easily and reused effectively, the accumulated information is essentially useless and the company's production costs will increase because similar knowledge must be recreated. Thus, successful knowledge management (KM) practices require an understanding of workers' information needs to ensure effective information-seeking activities when they perform long-term tasks.

Although some KMSs incorporate information retrieval (IR) functions, workers find it difficult to express their information needs by using short query terms (LaBrie & St. Louis, 2003; Pons-Porrata, Berlanga-Llavori, & Ruiz-Shulcloper, 2007; Ruthven, 2001). In many cases the worker may only have a general idea about a topic and may be uncertain about what information is required to execute the task at hand (Belkin, Oddy, & Brooks, 1982; Jansen, 2005; White, Jose, & Ruthven, 2003, White & Kelly, 2006). The anomalous state of knowledge (ASK) hypothesis, posits that a searcher's information needs arise from an anomaly in the state of knowledge; thus, there is a gap between their knowledge about a task and the perceived requirements of the task. The gap is called the *information need* and results in information-seeking activities to solve the problem, i.e., satisfy the searcher's information needs (Belkin et al., 1982; Byström & Järvelin, 1995; Mackay, 1960; Taylor, 1968; White et al., 2004). To address this problem, we propose an effective information-learning method based on a topic variation inspection process. The method considers an individual's search behavior pattern and the interests of workers' with similar information needs to learn the individual's dynamic information needs precisely in a timely manner. More specifically, we integrate the traditional implicit relevance feedback (IRF) algorithm (Kelly, 2004, Ruthven, Lalmas, & van Rijsbergen, 2003; White, 2004; White et al., 2004; Widyantoro, Loerger, & Yen, 2001) with a user information needs profiling process to improve the knowledge retrieval functions in KMSs. Conventional user profiling approaches only reflect the user's previous information needs based on their personal search behavior for

Received February 13, 2008; revised December 8, 2008; accepted July 21, 2009

• Published online 3 September 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21201

relevant documents. They do not consider possible changes in the topics searched by users with similar information needs. In this work we adopt the concept of collaborative filtering used in recommendation systems to show how other workers' experiences can enhance the target worker's collaborative search behavior patterns for the task at hand (Balabanovic & Shoham, 1997; Konstan et al., 1997).

Contemporary KMSs employ information technologies (IT), such as cooperative document management portals, groupware, and workflow management systems to facilitate access to knowledge assets, as well as the reuse and sharing of knowledge assets within and across organizations (Davenport & Prusak, 1998; Kankanhalli, Tanudidjaja, Sutanto, & Tan, 2003). A repository of structured explicit knowledge, especially in document form, is a codified strategy for managing knowledge (Davenport & Prusak, 1998; Gray, 2001). According to Gray (2001), codified knowledge helps knowledge workers exploit their organization's resources fully. Kankanhalli et al. (2003) observed that product-based firms, such as Xerox, Microsoft, and Hewlett-Packard, rely on both codification and sharing approaches to keep pace with dynamic and rapid changes in the business environment, i.e., the companies operate in a complex and high-volatility context. However, with the growing amount of information in organizational databases, KMSs face the increasingly difficult challenge of helping users find pertinent information efficiently. Thus, knowledge retrieval is considered a core component of systems that support workers engaged in knowledge-intensive tasks in a business environment (Abecker, Bernardi, Maus, Sintek, & Wenzel, 2000). To resolve the problem of retrieving needed information from a vast amount of codified knowledge, (IR) techniques coupled with workflow management systems (WfMS) are employed to support proactive delivery of task-specific information based on the context of the tasks within the overall process (Abecker et al., 2000; Fenstermacher, 2002). For example, the KnowMore system maintains task specifications (profiles) that detail the process-context of tasks and associated information items (Abecker et al., 2000). Context-aware, task-specific knowledge can thus be provided based on the task's specifications and the current execution context of the process. In another approach, a process meta-model that specifies the context of the objects is integrated with workflow systems to capture and retrieve information or codified knowledge within a process context (Kwan & Balasubramanian, 2003). The weakness of the above methods is that creating a task-based profile or specification requires human effort. Moreover, they employ push-based strategies that provide task-relevant information without considering the user's active search behavior. In other words, they cannot identify and track a worker's dynamic information needs (task needs) over time precisely. This is a critical issue because a worker's task-needs can emerge and change in different time frames during a task's execution.

It is widely agreed that information seeking is a difficult and complex process for workers during the execution of long-term projects/tasks (Kuhlthau, 1993; Spink, Wilson, Ellis, & Ford, 1998; Vakkari, Pennanen, & Serola, 2003). A number of studies on information search processes observe that users' information needs and search behavior patterns vary according to the problem stage they are in (Campbell & Van Rijsbergen, 1996; Ingwersen & Järvelin, 2005; Kuhlthau, 1993; Vakkari et al., 2003; Tang & Solomon, 1998). Kuhlthau's search process model differentiates a task into six stages with their associated characteristics. Specifically, it divides the information search process from the user's perspective into the following stages: task initiation, topic selection, prefocus exploration, focus formulation, information collection, and search closure. The objective is to observe how users locate and interpret information to form a perspective on a topic in different problem stages. During the search process, thoughts evolve from unclear and vague to clear, more focused understanding. The user's search behavior also changes with the formulation of a focus. Existing case studies showed that, in the early stages of a task's execution, students seek relevant information related to a general topic.(Kuhlthau, 1993; Vakkari et al., 2003). The user's search behavior also changes with the formulation of a focus. Therefore, it should be much easier to analyze users' dynamic information needs in terms of topic changes, rather than by analyzing changes in the keywords input to the system.

To put a worker's dynamic information needs into perspective, we propose a topic variation inspection model to facilitate the application of an IRF algorithm and collaborative filtering in user modeling. Relevance feedback (RF) improves the effectiveness of searches by reformulating or expanding the original query based on partial relevance judgments, i.e., feedback on part of the evaluation set (Rocchio, 1971; Salton & Buckley, 1990). By employing the implicit RF algorithm, the system can monitor users' access behavior unobtrusively to learn their information needs and modify their original queries. The method identifies changes in a worker's information needs by inferring those needs from documents the worker has browsed, read, or downloaded. Our approach bears some similarity to Campbell and Van Rijsbergen's (1996) Ostensive Model, which describes how users' information needs correspond to their knowledge states. The degree of uncertainty influences a user's perception of the "relevance" of information and results in different information-seeking activities. In addition, following previous studies (Campbell & Van Rijsbergen, 1996; Kuhlthau, 1993; Ruthven et al., 2003; Vakkari, 2000), we assume that a knowledge worker's uncertainty will decrease as the task progresses. This contrasts with the traditional relevance feedback model, which assumes that all information (i.e., information items that users regard as relevant) is generated by the same knowledge state. We consider that recently accessed documents reflect a worker's current task needs more accurately than those documents accessed earlier. Thus, a time factor is incorporated into the adapted IRF algorithm to reflect the relevance of the current information. In addition, we try to analyze changes in the worker's topic-needs based on the task's performance over time. For example, if a researcher is seeking relevant knowledge documents for a project, the research topics may vary as follows: "Event detection" => "Mining event change" => "Mining Patent change" => "Patent Mining" where the symbol "=>" indicates that the event on the left-hand side occurs before the event on the right-hand side. Therefore, we propose a learning model of information needs that can track a worker's topic-needs in different time frames during a task's performance. Because each topic in the topic taxonomy is associated with a corpus, we can compile and rank a set of task-relevant topics by calculating the similarity between the corpus of each topic and the worker's current task profile. The analysis results can be incorporated into the profile adaptation process for query expansion based on the RF algorithm (Kelly, 2004; Rocchio, 1966, 1971; Salton & Buckley, 1990) by means of the domain corpus. Moreover, to determine a worker's task needs, we try to predict changes in the worker's information needs by identifying other workers with similar information needs. That is, we consider possible changes in information needs in terms of how other workers' experiences could enhance the target worker's search results and satisfy their information needs for the task at hand. In a recent study Zhou, Ji, Zha, and Giles (2006) suggested that social actors can influence the evolutionary trends of topics, i.e., the transition from one topic to another. Accordingly, we propose an information needs learning method based on a collaborative topic variation inspection process. The method adjusts a worker's task profile according to their knowledge-seeking activities (e.g., implicit document access behavior) so that other workers with similar variations in topic needs over time can be identified. The proposed model not only identifies variations in a worker's task-needs for topics (i.e., self-topic needs) over time, but also analyzes the collective (i.e., collaborative actors') topic variation patterns and adjusts the worker's profile accordingly. Then, to satisfy the worker's dynamic task needs, codified knowledge relevant to the current task can be retrieved based on the adjusted task profile.

The remainder of this paper is organized as follows. Literature Review provides a review of related works. In Overview of the Dynamic Information Needs Learning Approach we formulate the problem and discuss our methodology. Measuring Variations in Topic-Needs over Time describes how we use the proposed model to measure variations in topic needs over time. The Topic Variation Inspection Process section explains how we use a personal topic variation inspection process and a collaborative topic inspection process to predict a worker's task-needs. We also present an algorithm for identifying the task needs of similar social actors via a topic variation matrix that is also presented. The experiment design and experiment results are presented in the next two sections, followed by Discussion and Implications, then Conclusion and Future Work.

Literature Review

Relevance Feedback in a Vector Space Model

Relevance feedback improves the search effectiveness of the automatic query reformulation process (Rocchio, 1966, 1971). The literature on information retrieval shows that relevance feedback applied in a vector model is an effective technique in a retrieval environment (Rocchio, 1966; Salton & Buckley, 1990). In a vector space model the key contents of a codified knowledge item (document) can be represented as a term vector (i.e., a feature vector of weighted terms) in an n-dimensional space, using a term-weighting approach that considers the term frequency, inverse document frequency, and normalization factors (Salton & Buckley, 1988). The *term transformation* steps, namely, case folding, stemming, and stop word removal, are performed during text preprocessing (Porter, 1980; Salton & Buckley, 1988; Witten, Moffat, & Bell, 1999). Then, *term weighting* is applied to extract the most discriminating terms (Baeza-Yates & Ribeiro-Neto, 1999). Let d be a codified knowledge item (document), and let $\vec{d} = <w(k_1, d), w(k_2, d), \ldots, w(k_n, d)>$ be the term vector of d, where $w(k_i, d)$ is the weight of a term k_i that occurs in d. Note that the weight of a term represents its degree of importance in representing the document (codified knowledge). The well-known *tf-idf* approach, which is often used for term (keyword) weighting, assumes that terms with a higher frequency in one document and a lower frequency in other documents are better discriminators for representing the first document. Let the term frequency $tf(k_i,d)$ be the occurrence frequency of term k_i in d, and let the document frequency $df(k_i)$ represent the number of documents that contain k_i. The importance of k_i is proportional to the term frequency and inversely proportional to the document frequency, which is expressed as follows:

$$w(k_i, d) = \frac{1}{\sqrt{\sum_i \left(tf(k_i, d) \times \log(N/df(k_i) + 1)\right)^2}} tf(k_i, d) \times \left(\log \frac{N}{df(k_i)} + 1\right), \qquad (2.1)$$

where N is the total number of documents. Note that the denominator on the right-hand side of the equation is a normalization factor that normalizes the weight of a term. The classic relevance feedback method proposed by Rocchio (1971) and the Ide_Dec_Hi (1971) method, which use a vector space model to derive the modified query $\vec{\mathbf{q}}_m$, are formulated in Equations 2.2 and 2.3, respectively (Baeza-Yates & Ribeiro-Neto, 1999):

$$\text{Standard\_Rocchio: } \vec{\mathbf{q}}_m = \alpha\vec{\mathbf{q}} + \beta\frac{1}{|D_r|}\sum_{\forall \vec{\mathbf{d}}_j \in D_r} \vec{\mathbf{d}}_j - \gamma\frac{1}{|D_n|}\sum_{\forall \vec{\mathbf{d}}_j \in D_n} \vec{\mathbf{d}}_j \qquad (2.2)$$

$$\text{Ide\_Dec\_Hi: } \vec{\mathbf{q}}_m = \alpha\vec{\mathbf{q}} + \beta\sum_{\forall \vec{\mathbf{d}}_j \in D_r} \vec{\mathbf{d}}_j - \gamma \max_{\textit{non-relevant}}(\vec{\mathbf{d}}_j) \qquad (2.3)$$

where D_r is the set of relevant documents and D_n is the set of non relevant documents, both of which are determined by the user; $|D_r|$ and $|D_n|$ denote the number of documents in the sets D_r and D, respectively; and α, β, γ are tuning constants. In Equation 2.3, $\max_{non\text{-}relevant}(\vec{\mathbf{d}}_j)$ denotes the highest non-relevant document. The two methods yield similar results (Baeza-Yates & Ribeiro-Neto, 1999). The Rocchio method sets $\alpha = 1$, Ide sets $\alpha = \beta = \gamma = 1$, and Harman (1992) sets $\beta = 0.75$, $\gamma = 0.25$. Salton and Buckley (1990) showed that the Ide_Dec_Hi algorithm performs slightly better than the Rocchio algorithm; hence, we adopted Ide_Dec_Hi as our baseline method.

Techniques for Modeling Users' Information Needs

Textual data, such as articles, reports, manuals, and know-how documents, are treated as valuable and explicit knowledge in organizations (Nonaka, 1994). Therefore, the first step of knowledge management usually involves the application of information retrieval (IR) and information filtering (IF) techniques to solve document management problems. In fact, IR and IF are considered core technologies that help organizations collect and process documents, mitigate the problem of information overload, and provide relevant information for knowledge workers to accomplish their tasks

IF systems are similar to conventional IR systems; however, rather than focus on supporting users' short-term information needs, IF systems emphasize the concept of personalization to support users' long-term information needs (Baeza-Yates & Ribeiro-Neto, 1999; Belkin & Croft, 1992; Elovici, Shapira, & Kantor, 2006; Shapira, Shoval, & Hanani, 1999). Basically, IF systems maintain user profiles and relevant information is delivered to users based on their profile. Thus, learning and maintaining users' profiles are important aspects of supporting long-term information services. Various approaches for learning users' interests or preferences from textual documents or Webpages have been proposed (Balabanovic & Shoham, 1997; Pazzani & Billsus, 1997; Middleton, Shadbolt, & Roure, 2004; Mostafa, Mukhopadhyay, Lam, & Palakal, 1997). Well-known approaches in information retrieval and information theory, such as Rocchio's algorithm, information gain theory, and the Bayesian classifier, have been modified and used to model or capture a user's dynamically changing interests. Most approaches require users' relevance feedback (RF) information, as well as explicit feedback (users' ratings on information items) or implicit feedback (users' access behavior), to achieve this goal. RF improves search effectiveness through query reformulation. Kelly and Fu's (2007) study of RF shows that determining a user's information needs based on their feedback can improve retrieval performance significantly. Moreover, various studies have demonstrated that applying RF in a vector model enhances IF (Middleton et al., 2004; Salton & Buckley, 1990; Widyantoro & Yen, 2005; Yang, Yoo, Zhang, & Kisiel, 2005).

Widyantoro et al. (2001) use a three-descriptor model to learn a user's multiple interests. The approach maintains a long-term descriptor to capture the user's general interests and a short-term descriptor to keep track of their more recent interests, which change more rapidly. An automatic weight-adjustment mechanism adjusts the weight of positive and negative descriptors to ensure that the short-term descriptor records major changes in the user's interests immediately. In addition, a profiling approach has been adopted in the workplace proposed to enhance knowledge retrieval and promote knowledge sharing among project-based or interest-based groups (Abecker et al., 2000; Agostini, Albolino, De Michelis, De Paoli, & Dondi, 2003; Davies, Duke, & Stonkus, 2003); and a cooperative agent architecture has also been developed to facilitate task-based information filtering within a work process (De Bra, Houben, & Dignum, 1997). In this the latter type of framework, information filtering techniques combined with an intelligent agent-based architecture are commonly adopted to streamline the delivery of knowledge from internal or external knowledge repositories (Spies, Clayton, & Noormohammadian, 2005; Ye & Fischer, 2002).

In recent years, several studies have stressed the importance of modeling users' interests or information needs for a specific work task incrementally in terms of topics, instead of as a set of weighted keywords or meta-data. Sieg, Mobasher, and Burke (2004) integrate user profiles and concept hierarchies to infer users' information contexts in order to enhance the original queries. In this way, IF systems learn users' current information needs from the RF and update the model for subsequent information filtering. Godoy and Amandi (2006) proposed an incremental concept clustering algorithm called WebDCC, which uses intelligent agents to build a profile of the user's search behavior for Web documents, i.e., it models a user's preferences and interests based on observations of their behavior. Learning approaches can maintain users' profiles adequately once the system receives feedback or observes changes in search behavior patterns; hence, such approaches are regarded as incremental learning techniques. In addition to modeling user's interests in terms of topics, incorporating the domain ontology (i.e., a hierarchical structure of domain topics/categories) into the profiling process is an effective way of modeling users' information needs for tasks (Godoy, Schiaffino, & Amandi, 2004; Middleton et al., 2004; O'Leary, 1988; Pons-Porrata et al., 2007).

Overview of the Dynamic Information Needs Learning Approach

Motivation

As mentioned above, we propose an information needs learning model based on a collaborative topic variation inspection process to track and analyze a worker's information needs for a task, i.e., task-needs. In this subsection we formulate the problem and basic concepts as follows.

- First, we focus on a worker's search behavior during the execution of knowledge-intensive tasks (tasks for short), such as research projects in academic institutions and product development tasks in R&D departments. The 2004 International Data Corporation (IDC) Report (Feldman, 2004) estimated that knowledge workers spend 15%–35% of their time just searching for information; however, on average, they succeed in finding the desired information less than 50% of the time. Nevertheless, information seeking or searching is regarded as a key activity of knowledge workers during the execution of tasks. Hence, knowledge (information) retrieval is a core component of KMSs.

- Generally, a worker's search behavior results from the fact that there is a gap between their knowledge about the task at hand and the perceived requirements of the task. Taylor (1968) described the continuous mental development of a user's information need as evolving from an "unconscious need" over a "conscious need" to a "compromised need." Subsequently, Belkin et al. (1982) extended the theory and put forward the hypothesis of the anomalous state of knowledge (ASK). The ASK hypothesis posits that a searcher's information need arises from an anomaly in the state of knowledge, such that there is a gap between their knowledge about a task and the perceived requirements of the task. The gap is called the *information need* and results in information-seeking activities to solve the problem (Belkin et al., 1982; Mackay, 1960; Taylor, 1968; White et al., 2004). Recall that, during the search process, the worker's thoughts evolve from unclear and vague to a clear, more focused understanding (Ingwersen & Järvelin, 2005; Kuhlthau, 1993; Vakkari et al., 2003). Therefore, it should be much easier to analyze users' dynamic information needs in terms of topic changes, rather than by analyzing changes in the keywords input to the system.
- Third, following Campbell and Van Rijsbergen (1996), who proposed the Ostensive Model to describe how users' information needs correspond to their knowledge states, we assume that the worker's uncertainty will decrease as the performance of the task progresses. We consider that recently accessed documents reflect a worker's current task needs more accurately than documents accessed earlier. Thus, a time factor, i.e., a decay function, is incorporated into the topic variation inspection process to identify the user's emerging or declining interest in work-task topics at different times.
- Finally, to support the execution of the current task, a worker usually needs to reference previously executed topics (the executed-task set) in the task-based domain ontology (DO). In this context, if the task profiles of two workers' have some degree of similarity and they exhibit similar rates of topic change in the constructed task-based DO, it is reasonable to assume that they will have similar information needs for tasks in the near future. Thus, we propose a collaborative topic variation inspection approach to predict a worker's information needs based on variations in the topic needs of similar workers over time. The idea is similar to that of collaborative filtering techniques used in recommender systems (Balabanovic & Shoham, 1997). Basically, such techniques identify users whose profiles are similar to the profile of the target worker. Then they provide recommendations or predictions based on the other workers' experiences to improve the incomplete content-based approach (Balabanovic & Shoham, 1997; Konstan et al., 1997). Therefore, we propose a topic variation inspection model to facilitate the application of an implicit relevance feedback (IRF) algorithm and collaborative filtering in user modeling.

Note that, when modeling users' task needs, we use a task-based topic taxonomy to conceptualize domain information about organizational activities. The topics and their corresponding topic profiles are used as references to adjust the task profiles based on their relevance (similarity) to the documents accessed by workers. Modeling worker's dynamic task-needs is performed in two phases: a personal profile adaptation phase and collaborative adaptation phase. Technically, the adaptation process takes the time factor into consideration. The rationale for our approach is that the more recently a document has been accessed, the more important it should be in reflecting a worker's current task needs. Thus, the contribution of the user's previous task needs (profile) to adjusting the task profile should be reduced according to the amount of time that has passed. The personal task profile only reflects previous information needs; that is, it does not consider possible changes in the topic-needs (i.e., information needs for tasks). Thus, we try to measure changes in a worker's information needs by identifying other workers with similar information needs, i.e., profile adaptation via collaborative filtering techniques. The rationale behind collaborative profile adaptation is that workers with similar changes in previous information needs are likely to have similar changes in future information needs; thus, possible changes in a worker's information needs can be inferred from changes in the information needs of similar workers.

The Proposed Approach

The personal topic variation inspection phase adjusts task profiles incrementally based on the documents accessed by the worker over time. That is, the documents accessed most recently are deemed more important than those accessed in the early stages of the task's execution. The collaborative topic variation inspection phase uses variations in the topic patterns of similar workers to predict a target worker's future task-needs and adjust their task profile accordingly. We also use an event-based technique to trigger the profile adaptation step based on the results of the topic variation process. An event occurs when a worker accesses a document at a specific time, and each event-based topic need is modeled as a weighted topic. The weight of a topic is derived by considering the similarity between the topic profile and the profile of the document accessed by the worker at that time. Variations in a worker's topic needs can be measured by the difference the topic's weights at the two timepoints. Figure 1 provides an overview of the proposed methodology.

Phase 1: Personal topic variation inspection process. When a worker accesses a document, the system captures information about their search behavior. The profile adaptation phase considers the effect of the time factor and the worker's behavior (the document accessed) in order to adjust the corresponding task profile with the help of a task-based topic taxonomy. Note that other workers' feedback is not considered at this point. Since a current task is often related to some previously executed tasks in the organization, task-based topics play an important role as references that can provide workers with task-relevant documents (Middleton et al., 2004; O'Leary, 1988). In a previous work (Liu & Wu, 2008), we showed that relevant topics are points of reference

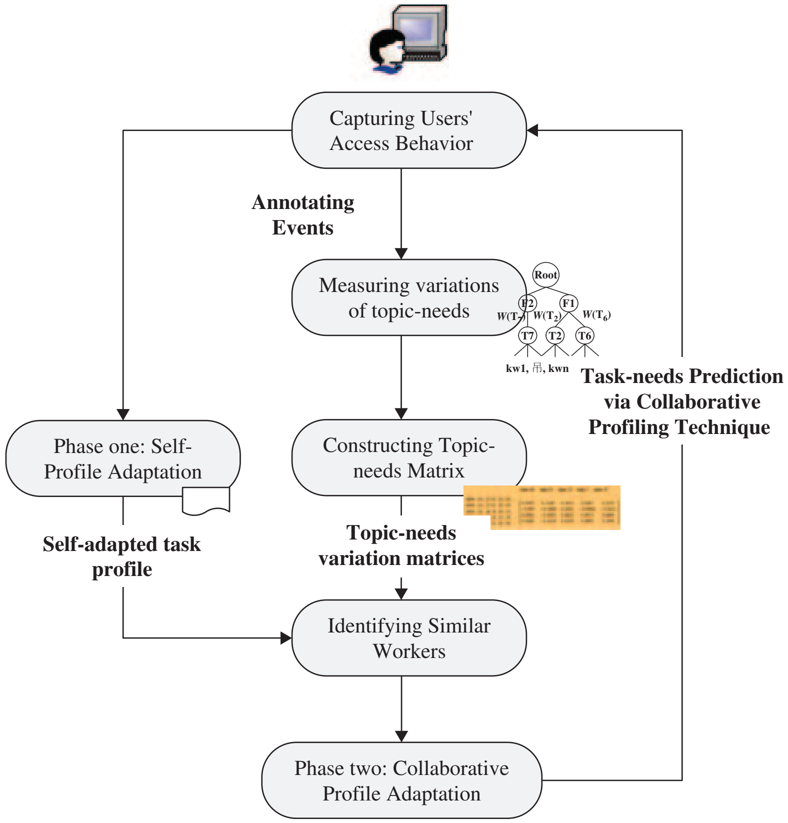

Capturing Users' Access Behavior
Annotating Events
Measuring variations of topic-needs
Root
F2
F1
W(T7)
W(T2)
W(T6)
T7
T2
T6
kw1, 吊, kwn
Task-needs Prediction via Collaborative Profiling Technique
Phase one: Self-Profile Adaptation
Constructing Topic-needs Matrix
Self-adapted task profile
Topic-needs variation matrices
Identifying Similar Workers
Phase two: Collaborative Profile Adaptation

FIG. 1. Overview of the approach.

that are very helpful for generating task profiles, especially when only a few documents are accessed in the early phase of a task's execution. The topics and their corresponding topic profiles are used as references to adjust task profiles for generating self-adapted profiles according to their relevance (similarity) to the documents accessed by workers. Details of the proposed personal topic variation inspection process are given in the first subsection of The Topic Variation Inspection Process (below).

Phase 2: Collaborative actors' topic variation inspection process. Instead of specifying changes in workers' information needs explicitly, we try to capture variations in their needs through the documents they access. To this end, we use variations in workers' topic needs over time to model and predict the target worker's possible task-needs. Variations in each worker's topic-needs over time are expressed as a topic-needs variation matrix, i.e., a time-period by topic-needs matrix. A similar approach is used to find workers with similar variations in topic needs based on the derived topic-variation matrix and personal profiles in a time window. After workers with similar needs have been identified, their task needs (at time $T+1$) are used to predict the worker's task needs at time $T+1$, which are modeled as a collaborative profile. The derived collaborative profile is then combined with the self-adapted profile to generate a new task profile that represents the target worker's future task needs at time $T+1$. The proposed collaborative topic variation inspection process is described in detail in the second subsection of The Topic Variation Inspection Process (below).

Notations

The notations used in this work are defined in Table 1.

Measuring Variations in Topic-Needs Over Time

In this section we describe the proposed approach for measuring the variations in a worker's topic needs over time, i.e., the information needs reflected in the topic taxonomy.

Capturing Workers' Relevance Feedback Behavior

When a worker accesses a document, the proposed system stores the information about the corresponding relevance feedback (judgment) behavior. Information about a worker's feedback behavior, i.e., the behavior related to documents, is gathered by our system's online user behavior tracker. The following example explains how the system captures and stores a worker's access behavior patterns.

Example: Assume that a worker, "Steven," searches for documents in the K-Support system, and finds a document entitled "Learning User Interest Dynamics with a Three-Descriptor Representation" that may help him with his current task. Steven "reads" the document on "2005-10-31" at "21:05:03" and the system records this information.

DOI: 10.1002/asi

TABLE 1. Notations used in this work.

Notation	Definition	Section
T	The index of the time that the latest event occurred	Section Measuring Variations of a Specific Topic
$TW_{t,T}$	The time weight of the event that occurred at time t with respect to time T	Section Measuring Variations of a Specific Topic
NW_t^i	The topic-need weight of topic i for the event that occurred at time t	Section Measuring Variations of a Specific Topic
$NV_{d,e}^i$	The variation in the worker's information needs for a specific topic i between time e and time $d(d< e)$	Section Measuring Variations of a Specific Topic
$\overrightarrow{profile}_{T+1}$	The personal profile generated at time T	Section Phase One: Personal Topic Variation Inspection Process
u_a/u_x	The target worker / similar worker	Section Phase Two: Collaborative Topic Variation Inspection Process
$Sim_T(u_a, u_x)$	The similarity between the target worker u_a and the worker u_x at time T, obtained from the *SimScore*(u_x)	Section Phase Two: Collaborative Topic Variation Inspection Process
δ_V/δ_D	The tuning parameters used to adjust the relative weights of the self-adapted profile and the collaborative profile	Section Information Needs Modeling Based on the Topic Variation Inspection Process
T'_{ux}	The latest time index of the similar candidate variation matrix $WM^{T'}(u_x)$ of worker u_x	Section Information Needs Modeling Based on the Topic Variation Inspection Process
$NV^i_{T'_{ux},T'_{ux}+1}(u_x)$	The variation degree of topic i between time T'_{ux} and time $T'_{ux}+1$ of u_x	Section Information Needs Modeling Based on the Topic Variation Inspection Process
$\overrightarrow{doc}_{T'_{ux}+1}(u_x)$	The document profile of a document, *doc,* accessed at time $T'_{ux}+1$ by the worker u_x	Sections Phase One: Personal Topic Variation Inspection Process & Information Needs Modeling Based on the Topic Variation Inspection Process
$\overrightarrow{topic}_i$	The topic profile of topic i in the topic taxonomy	Sections Phase One: Personal Topic Variation Inspection Process & Information Needs Modeling Based on the Topic Variation Inspection Process

In the above example, the stored information is {"Steven," "2005-10-31," "21:05:03," "reads," "Learning User Interest Dynamics with a Three-Descriptor Representation"}. Each attribute, except the time attribute, is converted into an identifiable number. Hereafter, we use *event* to denote the users' actions when they access a document. In this paper we only consider four kinds of events, namely, "downloading documents," "downloading reports of documents," "reading documents online," and "uploading documents." Workers upload documents that they regard as relevant or helpful to their research topics. They may also read and download documents or download notes about documents that are of personal interest.

Measuring Variations of a Specific Topic

We consider two factors when measuring the variation in information needs for a specific topic i in the topic taxonomy for each event that occurs at time t. The first is the time factor, called the *time weight* $TW_{i,T}$; and the second is the relevance degree of topic i, called the *topic-need weight* NW_t^i. In the following we explain the concepts of time weight and topic-need weight over time. Note that $TW_{i,T}$ is the time weight of an event that occurred at time t with respect to time T, as described in Equation 5.2; and NW_t^i is the topic-need weight of topic i for the event that occurred at time t. The latter is obtained by calculating the similarity (using a vector-based cosine method) between the document accessed at time t and the profile of topic i. When a worker accesses a document, the system calculates the topic-need weight for the corresponding event.

After obtaining the task-need weights for all of the worker's events, the variation in topic needs over time can be measured, as shown in Equation 4.1. Given two timepoints, d and e (where $d < e$), let $NV_{d,e}^i$ denote the variation in the worker's information needs for a specific topic i between time d and time e. The accumulated topic needs at time e equal the summation of $TW_{t,e} \times NW_t^i$ for $t = 1$ to e.

$$NV_{d,e}^i = \sum_{t=1}^{e} TW_{t,e} \times NW_t^i - \sum_{t=1}^{d} TW_{t,d} \times NW_t^i \qquad (4.1)$$

The relevance degree of topic i is different at time d and time e; thus, NW_t^i is exploited to take this factor into consideration. The time weight is used to reflect the effect of time decay on topic needs between time d and e. Since the measurement considers accumulated topic needs over time, the events that occurred before time d and time e are also considered.

A Representative Model

A representative model, i.e., a vector-based model, is defined to represent the variations in a worker's topic-needs over time. Such variations are expressed as a time-period in a topics matrix, i.e., a topic-needs variation matrix comprised of several topic-needs variation vectors. Let $NV_{d,e}$, defined in Equation 4.2, denote the variation vector of the worker's topic needs between time e and d. The measurement of the variation of a specific topic i, i.e., $NV_{d,e}^i$, is defined in Equation 4.1.

$$NV_{d,e} = <NV_{d,e}^1, NV_{d,e}^2, \ldots, NV_{d,e}^i, \ldots, NV_{d,e}^q> \qquad (4.2)$$

DOI: 10.1002/asi

$$\begin{array}{c|ccccc} & topic\ 1 & topic\ 2 & topic\ 3 & topic\ 4 & topic\ 5 \\ \hline \begin{array}{c} 2003-09-24\ 09:57:00 \\ 2003-10-07\ 18:25:42 \\ 2003-10-13\ 14:13:00 \\ 2003-10-13\ 14:18:00 \\ 2003-10-13\ 14:25:00 \\ 2003-10-14\ 14:49:30 \end{array} & \multicolumn{5}{c}{\begin{bmatrix} 0.066 & -0.013 & -0.024 & -0.066 & 0.065 \\ 0.447 & 0.328 & 0.014 & 0.074 & 0.078 \\ 0.06 & 0.026 & 0.031 & 0.111 & 0.080 \\ 0.04 & 0.015 & 0.036 & 0.109 & 0.044 \\ 0.05 & -0.019 & 0.019 & 0.033 & 0.064 \end{bmatrix}} \end{array}$$

FIG. 2. Example of a topic-needs variation matrix (*VM*).

Variations in topic needs between consecutive timepoints are expressed as a time-period by the topic matrix *VM*. An element $VM_{p,i}$ in the matrix represents the variation of topic i during time-period p (e.g., from time d to e). A row in the matrix, $VM[j]$, denotes a variation vector of topic needs. Figure 2 shows an example of a topic-needs variation matrix.

Example: The variation matrix shown in Figure 2 is a 5×5 matrix. The variations in topic needs represented by the matrix cover the period 2003-09-24 (09:57:00) to 2003-10-14 (14:49:30), and the value of each element in the matrix represents the variation of the corresponding topic. For example, the value 0.447 represents the variation in topic needs for topic 1 from 2003-10-07 (18:25:42) to 2003-10-13 (14:13:00). Let *t1* denote the timepoint 2003-09-24 09:57:00, and let *t2* denote the timepoint, 2003-10-07 18:25:42. In this case, $NV^1_{t1,t2} = 0.066$, which represents the variation in topic needs for topic 1 between *t1* and *t2*; and $NV_{t1,t2} = \langle 0.066, -0.013, -0.024, -0.066, 0.065 \rangle$, which represents the variations in topic-needs for all topics between *t1* to *t2*. The variations in topic needs over time are represented as set of topic-needs variation vectors.

The Topic Variation Inspection Process

In this section we describe the collaborative topic variation inspection process, which adjusts task profiles based on individual and collective search behavior. The two phases of profile adaptation, which are based on collaboration, are discussed in the following two subsections. In addition, we explain how we integrate the derived collaborative profile with the personal profile to predict the target worker's task needs.

Phase One: Personal Topic Variation Inspection Process

When the system detects an event related to a worker's access behavior, it captures and records the document accessed by the worker. The event triggers the personal profile adaptation process, which adjusts the worker's task profile based on the corresponding event. A modified IRF algorithm, adapted from the techniques applied in the Ide_Dec_Hi algorithm, is used to adjust the workers' task profiles. The proposed profiling technique is defined in Equations 5.1 and 5.2. Let T denote the timepoint that the worker accessed the last document; and let $\overrightarrow{profile}_{T+1}$ denote the worker's task profile generated at time T, which can be used to model their task-needs at time $T+1$.

$$\overrightarrow{profile}_{T+1} = \alpha \times Decay(\overrightarrow{profile}_T) + [\lambda \overrightarrow{topic}_T + (1-\lambda)\overrightarrow{doc}_T] \tag{5.1}$$

The task profile $\overrightarrow{profile}_{T+1}$ is generated from previous task profile $\overrightarrow{profile}_T$ applied with a decay function, and refined by using the current information needs derived from the document accessed at time T. The current information needs are divided into a document profile, $\overrightarrow{doc}_T$, and an aggregate topic profile, $\overrightarrow{topic}_T$. Intuitively, $\overrightarrow{doc}_T$ is the profile (feature vector) of the document accessed at time T. The relevance degree of a topic i to the document accessed at time T is obtained by calculating the similarity (cosine measure) between $\overrightarrow{topic}_i$ and $\overrightarrow{doc}_T$. The $\overrightarrow{topic}_T$ profile is derived from the topic profiles of relevant topics in the positive topic set, as well as nonrelevant topics in the negative topic set. We use a parameter, λ, to adjust the weights of the document profile and the aggregate topic profile, as shown in Figure 3.

$$\begin{aligned} Decay(\overrightarrow{profile}_T) &= \sum_{t=1}^{T-1} TW_{t,T} \times [\lambda \overrightarrow{topic}_t \\ &\quad + (1-\lambda)\overrightarrow{doc}_t] \quad \text{where} \\ \text{Time Weight: } TW_{t,T} &= \frac{\textit{the actual time for } t - ST}{\textit{the actual time for } T - ST} \end{aligned} \tag{5.2}$$

$Decay(\overrightarrow{profile}_T)$ represents the accumulated task needs from the beginning of the task to the current time T. Thus, in this work we incorporate the time decay function of the previous task-profile, as given in Equation 5.2, where $\overrightarrow{profile}_T$ denotes the previous task profile generated at time $T-1$, and represents the previous task needs. Specifically, $\overrightarrow{profile}_T$ is the aggregate of topic profiles and document profiles derived from the starting time ST to $T-1$. Generally, the more recently a document was accessed, the more important it should be in reflecting the worker's current task needs. $TW_{t,T}$ is the time weight of an event that occurred at time t with respect to T, and is defined as the ratio of the time difference $t-ST$ to $T-ST$. Thus, $Decay(\overrightarrow{profile}_T)$ reflects the effect of the previous task profile on the current task profile more accurately with TW than just using $\overrightarrow{profile}_T$. Accordingly, to learn the users' dynamic information needs we propose three

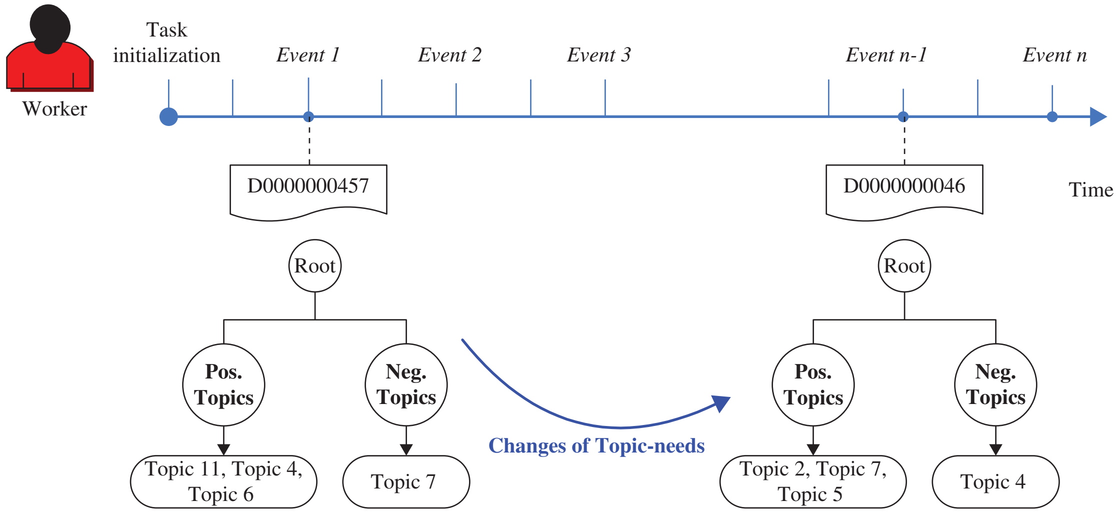


FIG. 3. Example of modeling a worker's task-needs.

self-profile adaptation methods that consider the topic variation factor and the time factor. We call the methods *P-Time, P-Topic*, and *P-Topic&Time*.

Phase Two: Collaborative Topic Variation Inspection Process

The variations in each worker's topics-needs can be represented by a topic-needs variation matrix (*VM*), as shown in Figure 2. Specifically, the variations in each worker's topic needs over time are expressed as a time-period by a topic matrix based on topics in the taxonomy, i.e., a topic-needs variation matrix comprised of several topic-needs variation vectors. As a result, workers with similar topic-needs can be identified by their topic-need variation matrices and personal profiles in a time window. That is, to identify similar workers, we consider workers with similar variations in topic needs and similar personal profiles simultaneously. Figure 4 shows the proposed algorithm for identifying workers with similar task needs. We explain the algorithm in detail in the following.

First, the variation matrix $VM^T(u_a)$ of the target worker is trimmed to a $w * q$ variation matrix $WM^T(u_a)$, which only contains the latest w variation vectors used to identify workers with similar variation matrices and task profiles. For each compared worker u_x, a sliding window W is used to locate the part of the variation matrix of u_x that is similar to $WM^T(u_a)$. $WM^{T'}(u_x)$ is the variation matrix of u_x generated according to $VM^T(u_x)$ and the sliding window W, and T' is the latest time index in the window W. Note that W denotes the sliding window whose size is w, where $w \leq \text{row}(VM^T(u_a))$, and q represents the number of topics in the topic taxonomy. An example of the latest w variation vectors of the trimmed variation matrix is shown in Figure 5. In this case, the size of the window, w, is equal to four. The proposed algorithm tries to capture variations in the target worker's topic-needs for the time period that is closest to the latest time index of the worker's search activities. We believe that some of the workers who perform similar tasks will have similar variations in topic-needs during that time period.

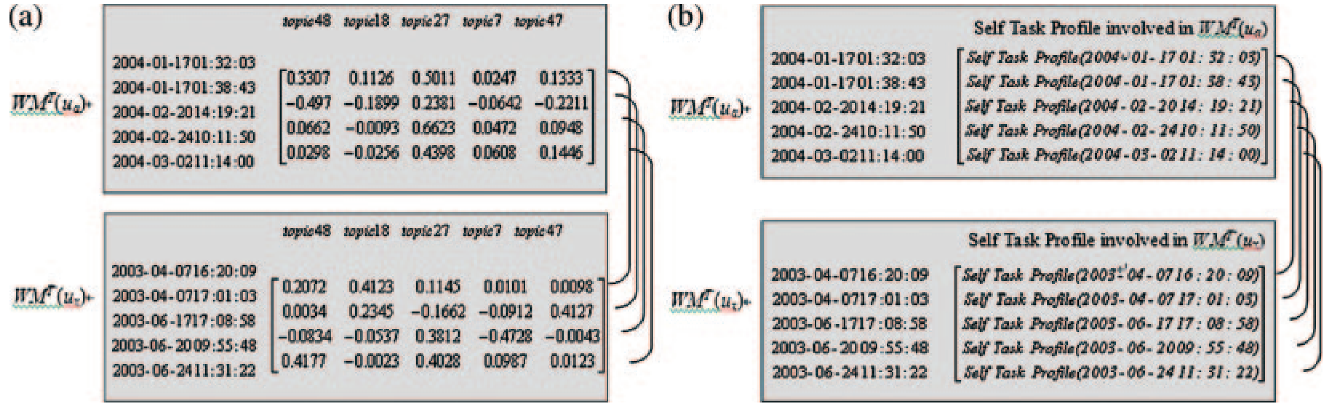
We employ a trimmed matrix instead of a complete variation matrix because it is not easy to find workers with similar changes in topic-needs for the whole task in the long term. In general, users only have similar topic-needs for a short period of time; therefore, we set a time window to make comparisons among the workers. In addition, the computational cost of comparing the target worker's matrix with those of the other workers would be prohibitive. In Figure 4, lines 3–21 describe the procedure for finding candidate workers with similar topic-needs based on the candidate variation matrix for each compared worker u_x. The candidate variation matrix with the highest similarity score among all the candidates of u_x is selected as the most similar variation matrix to that of u_x. The calculations (lines 16–19) of the similarity *SimScore* of u_x and u_a are divided into two parts: 1) calculation of the similarity *SimVM* based on the topic-needs variation vectors; and 2) calculation of the similarity *SimTP* based on the personal profiles. A parameter η is used to balance the relative importance of *SimVM* and *SimTP*. In our application, we set $\eta = 1/2$; that is, the similarity scores of the variation vectors and the personal profiles are equal. The workers with the top-N ranked similarity scores are selected as the similar workers of u_a. The value of N should be set according to the application domain. Figure 5a,b illustrate, respectively, the calculation of the similarity scores based on the variation matrices and task profiles in the sliding window.

Information Needs Modeling Based on the Topic Variation Inspection Process

After identifying workers similar to the target worker, we use their variation matrixes, i.e., similar candidate variation matrixes determined by the algorithm, can be used to predict the target worker's potential task needs, as shown in Equations 5.3 and 5.4.

```
Input:: $VM^T(u_a)$, $W$
Output:: SimilarWorkerList    // the list of similar workers
    function FindSimilarWorker($VM^T(u_a)$, $W$){
        Trim $VM^T(u_a)$ to a $w * q$ variation matrix $WM^T(u_a)$
        // which only keeps the last $w$ variation vectors;
        foreach compared worker $u_x${
            Set $SimScore(u_x) = 0$
            Slide the window $W$ on $VM^T(u_x)$ to derive $WM^{T'}(u_x)$
                from row 1 to row (row($M_y^T$)-w + 1), do {
                Let $WM^{T'}(u_x)$ be the variation matrix of $u_x$ generated based on $VM^T(u_x)$ and the sliding window $W$, when $W$ is
                moving on $VM^T(u_x)$; $T'$ is the latest time index in the window $W$
                  for the variation matrix $WM^{T'}(u_x)$ covered by $W$, do{
                    Set $SimVM = 0$, $SimTP = 0$
                    foreach variation vector $WM^{T'}(u_x)$[j] of $WM^{T'}(u_x)$ do {
                        Let $WM^T(u_a)$[j] be the corresponding variation vector of $M^T(u_a)$
                        $SimVM = SimVM + simlarity(WM^{T'}(u_x)[j], WM^T(u_a)[j])$
                    }
                    $SimVM = SimVM / w$
                    foreach personal profile $TP(u_x)$[k] involved in $WM^{T'}(u_x)$ do{
                        Let $TP(u_a)$[k] be the corresponding task profile of $WM^T(u_a)$
                        $SimTP = SimTP + simlarity(TP(u_x)[k], TP(u_a)[k])$
                    }
                    $SimTP = SimTP / (w + 1)$
                    if $((\eta * SimVM + (1 - \eta) * SimTP) > SimScore(u_x))$ then{
                        $SimScore(u_x) = \eta * SimVM + (1 - \eta) * SimTP$
                        Set $WM^{T'}(u_x)$ as the candidate (similar) variation matrix of $u_x$
                    }
                }
            }
        }
        Add the workers with top-N $SimScore$ to SimilarWorkerList;
        return SimilarWorkerList;
    }
```

FIG. 4. The algorithm for identifying similar workers.



$$\begin{bmatrix} 0.3307 & 0.1126 & 0.5011 & 0.0247 & 0.1333 \\ -0.497 & -0.1899 & 0.2381 & -0.0642 & -0.2211 \\ 0.0662 & -0.0093 & 0.6623 & 0.0472 & 0.0948 \\ 0.0298 & -0.0256 & 0.4398 & 0.0608 & 0.1446 \end{bmatrix}$$

$$\begin{bmatrix} 0.2072 & 0.4123 & 0.1145 & 0.0101 & 0.0098 \\ 0.0034 & 0.2345 & -0.1662 & -0.0912 & 0.4127 \\ -0.0834 & -0.0537 & 0.3812 & -0.4728 & -0.0043 \\ 0.4177 & -0.0023 & 0.4028 & 0.0987 & 0.0123 \end{bmatrix}$$

$$\begin{bmatrix} Self\ Task\ Profile(2004-01-1701:32:03) \\ Self\ Task\ Profile(2004-01-1701:38:43) \\ Self\ Task\ Profile(2004-02-2014:19:21) \\ Self\ Task\ Profile(2004-02-2410:11:50) \\ Self\ Task\ Profile(2004-03-0211:14:00) \end{bmatrix}$$

$$\begin{bmatrix} Self\ Task\ Profile(2003-04-0716:20:09) \\ Self\ Task\ Profile(2003-04-0717:01:03) \\ Self\ Task\ Profile(2003-06-1717:08:58) \\ Self\ Task\ Profile(2003-06-2009:55:48) \\ Self\ Task\ Profile(2003-06-2411:31:22) \end{bmatrix}$$

FIG. 5. (a) Example of calculating *SIMV* (average value of the similarity of variation vectors). (b) Example of calculating *SIMTP* (average value of the similarity of self-task profiles).

Note that time T'_{ux} is the latest timepoint in the presented candidate variation matrix. The variation vectors immediately after time T'_{ux} in the matrix can be regarded as the possible changes in topic needs that the target worker will experience in the near future. We propose two approaches for predicting the worker's potential task needs based on the behavior patterns of similar workers at time $T'_{ux} + 1$. The first, called the Coll_Topic Variation approach, is based on the variation in topic needs. It generates a collaborative profile in which variations in the topic needs of similar workers from time T'_{ux} to $T'_{ux} + 1$ are used to predict possible variations in the target worker's topic needs at time $T + 1$. The second method, called the Coll_Document method, is based on the documents accessed at time $T'_{ux} + 1$. In this case, the documents that similar workers accessed at time $T'_{ux}+1$ are used to derive the collaborative profile. The linear combination approach detailed below is used to integrate the derived collaborative profile with the personal profile to generate a new task profile $\overrightarrow{coll\_profile}_{T+1}$, which represents the target worker's future task needs.

Coll_Topic Variation Method (5.3)

$$\times \frac{\sum_{u_x \in Similar\ Worker\ Set\ of\ u_a} \left(Sim_T(u_a, u_x) \times \sum_{i=1}^{q} NV^{i}_{T'_{ux}, T'_{ux}+1}(u_x) \times \overrightarrow{topic_i}\right)}{\sum_{u_x \in Similar\ Worker\ Set\ of\ u_a} Sim_T(u_a, u_x)}$$

Coll_Document Method (5.4)

$$\times \frac{\sum_{u_x \in Similar\ Worker\ Set\ of\ u_a} \left(Sim_T(u_a, u_x) \times \overrightarrow{doc}_{T'_{ux}+1}(u_x)\right)}{\sum_{u_x \in Similar\ Worker\ Set\ of\ u_a} Sim_T(u_a, u_x)}$$

In the Coll_Topic Variation method, the predicted profile of the target worker is the weighted combination of the personal profile, i.e., $\overrightarrow{profile}_{T+1}$ in Equation 5.1 and the accumulated collaborative topic-variation profile derived from variations in the topic needs of similar workers. Here, we use a parameter δ_V/δ_D to adjust the relative weights of the self-adapted personal profile and the collaborative profile. Each worker's topic-variation profile is obtained by multiplying the profiles of topics by the corresponding variation degrees $NV^{i}_{T'_{ux}, T'_{ux}+1}(u_x)$. The individual topic variation profile represents the variation in the topic needs of the corresponding worker; while the collaborative topic-variation profile represents similar workers' weighted topic-variation profiles in terms of their similarity to that of the target worker. The difference between the Coll_Document method and the Coll_Topic Variation method is that the documents accessed at time $T'_{ux}+1$ by similar workers are used to derive the collaborative document profile.

Experiments

We conducted three experiments to evaluate the effectiveness of the proposed information needs leaning methods. The first subsection provides an overview of the K-Support system and the experiment setup; the second subsection describes the evaluation metrics; and the third subsection details the experiment procedure and methods.

Experiment Setup

We conducted experiments in a real application domain used for research tasks in a research institute's laboratory. For this investigation context we developed a task-based K-Support portal to deliver relevant documents based on the user's work task and their current information needs. We describe the proposed "K-Support" system below.

Overview of the K-Support system. In task-based environments, codified knowledge and human resources are important knowledge assets that can be used to accomplish organizational tasks. The K-Support portal is a Web-based application that allows workers to retrieve, organize, and share task-relevant documents (Liu & Wu, 2008). The system architecture comprises four implementation layers: knowledge resource collection, knowledge acquisition, knowledge modeling, and a Web-based front-end application. A user can log into the system to search for or share a task-relevant document. In the knowledge acquisition layer, the user behavior tracker and log-parsing engine analyze log-files to track the user's interaction with the system. We do not ask users to provide feedback for every document. The system can monitor the user's search behavior and the user's document feedback behavior patterns are gathered by the back-end user behavior tracker. That is, we observe the users' natural browsing, reading, and access behavior patterns instead of instructing them to follow specific steps in the proposed portal. More details about the task-based K-Support system can be found in the above-mentioned work.

To evaluate the approach proposed in this paper, we use a system with four modules: a user behavior tracker, a domain topic taxonomy, an information-needs variation inspector, and a profile handler. The developed system is an extension of the framework of our previously proposed K-Support system, i.e., we have added the information-needs variation inspector module. To learn a worker's dynamic information needs, the information-needs variation inspector is designed to fully utilize the information about the work context, i.e., the search behavior of similar workers and changes in the domain topics. The framework can be integrated with a KMS or project management system to design the information retrieval function. In addition, the proposed approach can be generalized via the presented framework to support workers' information seeking and retrieval activities when executing knowledge-intensive tasks. In the following, we discuss the study setting, dataset, and evaluation metrics.

Study setting and dataset. This work extends the previous framework to improve the most important functions, i.e., the knowledge retrieval functions, based on the user's work task and their current information needs. We chose tasks performed in the Department of Information Management of a major Taiwanese university for evaluation. The tasks included system development, thesis writing, and project surveys, all of which can be regarded as knowledge-intensive tasks. The subjects were graduate students who were engaged in different tasks. Four research issues were selected as the evaluation targets, namely, information technology service management, patent analysis for business intelligence, product recommendations, and knowledge management systems. The students needed to access documents for a specific task in the proposed digital workspace for use in a regular weekly meeting held in the research institute. As the subjects needed to upload between one and three documents that

DOI: 10.1002/asi

were relevant to their research every week, we assumed that the system could track changes in each worker's topics of interest during the task's performance. Each research project issue covered several related research topics. Although our evaluation tasks were implemented in the same department, they belonged to different projects with their own research topics. Twelve subjects were selected as test workers for the evaluation. The evaluation period for each target task was determined by experts who evaluated the proposed methods. We sampled the evaluation subjects based on certain criteria, one of which was the problem stage of the long-term task that the worker was in. As each subject was in a different problem stage (i.e., cognitive state), the size of the dataset and the number of participants were restricted in the experiments. Basically, we followed Kuhlthau's (1993) information search process model (ISP model) and Vakkari et al.'s task-based information-seeking theories (2000, 2003). Some of the empirical longitudinal studies conducted by these authors show that users' information needs vary in different stages. This factor motivated the current research, as mentioned in the Introduction. In this work we divide a user's search process into three stages: the pre-focus, focus formulation, and post-focus stages. Knowing the worker's current problem stage while conducting the experiments can help us explain the results of topic variation, which are detailed in Experiment Results (below).

Evaluation Metrics

The goal of the experiments is to evaluate the performance of the proposed information needs learning model via the topic variation inspection process. The IR evaluation methodology focuses on the evaluation of quantitative or qualitative data (Chen, Fan, Chau, & Zeng, 2001). Retrieval effectiveness is the most commonly used criterion for quantitative evaluation, and the effectiveness of information retrieval is normally measured by the precision and recall rates (Chen et al., 2001; Croft, 1995; Salton & McGill, 1983). On the other hand, qualitative evaluation of an IR system can be based on the analysis of questionnaires that request information about various evaluation items, such as user satisfaction, usability, and learning ability. Qualitative evaluation is much more suitable for evaluating the effectiveness of users' interactive search activities. Our evaluation method focuses on the retrieval results. We compare the performance of various methods in terms of their retrieval effectiveness. Specifically, we use the precision rate, recall rate, and F-measure to compare the methods (Rijsbergen, 1979; Salton & McGill, 1983; Witten et al., 1999).

Precision and recall. Precision is the fraction of retrieved documents that are relevant, while recall is the fraction of known relevant documents that are retrieved. To calculate the precision and recall rates, we asked domain experts and experienced workers to manually label documents that were highly relevant for each task. Although this is very time-consuming, it ensures the quality of our answer set for each evaluation task. The precision rate for an evaluation task e_r is the ratio of the total number of relevant documents retrieved to the number of top-N support documents in the presented system. The recall rate for an evaluation task e_r is the ratio of the total number of relevant documents retrieved to the total number of relevant documents specified by experts.

F-measure. To assess the relative importance of the precision and recall rates, a combination metric, the F-measure (Rijsbergen, 1979; Witten et al., 1999), is used to adjust the relative weights of precision and recall to find a trade-off between the two metrics. The function of β is to adjust the importance of the recall rate relative to that of the precision rate. If $\beta = 0$, F_β coincides with precision, and if $\beta = \infty$, F_β coincides with recall. To compare the methods in this experiment, we consider the precision ($\beta = 0$), recall (where $\beta = ^\infty$), and F-measure ($\beta = 0.5$), i.e., precision is more important than recall.

$$F - measure = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall} \qquad (6.1)$$

In this work, finding relevant documents based on a few retrieval results is much more important than finding all relevant documents. Therefore, we emphasize precision more than recall because we want to determine which method is better able to reject extraneous documents rather than find all relevant documents (Salton & Buckley, 1988). Moreover, the higher precision rate (i.e., $\beta = 0$ in the F-measure) reflected in the experiment results shows that the proposed methods are suitable for interactive work environments where workers are under pressure to find task-relevant documents and do not have time to review a large number of retrieved documents.

Experiment Procedure and Methods

As mentioned above and illustrated in Figure 1, the model uses an event-based technique (i.e., an event occurs when a worker accesses a document at a specific time) to trigger the information needs learning and profile adaptation process. Recall that we did not ask the subjects to provide explicit feedback on the documents. The system can monitor and record a user's search behavior with the back-end user behavior tracker. When a user logs into the system for a specific work task, four types of event are recorded, namely, "download document," "download reports of documents," "read document online," and "upload document." With the proposed topic variation inspection method, the system can deliver task-relevant documents based on the learning results.

The objective of the three experiments was to compare the effectiveness of the proposed adaptive information needs learning methods, namely, the P-Time, P-Topic, P-Topic&Time, Coll_Topic Variation, and Coll_Document methods, with that of the baseline method, S_P (primitive self-profiling). Details of each method are given in Table 2.

In Experiment 1 we evaluate the effectiveness of the baseline method and three of the proposed methods for profile adaptation via the personal topic variation inspection process, as discussed in Phase 1 of the proposed methodology (see above). The baseline method is the traditional relevance feedback (RF) technique, which is widely used in IF studies to determine users' dynamic interests, preferences, and information needs. Most studies in this area adopt the Rocchio method as the baseline to compare the performance of a proposed method (Salton & Buckley, 1990; Widyantoro & Yen, 2005; Yang et al., 2005). We refer to the classical relevance feedback method proposed by Rocchio (1971) and the Ide method (1971) for query reformulation, as formulated in Equations 2.2 and 2.3. The S_P method is similar to the Rocchio method, except that the nonrelevant feedback part of the equation is removed because most studies suggest that information about relevant documents is more important than the content of nonrelevant documents (Salton & McGill, 1983; Salton & Buckley, 1990; Liu & Wu, 2008).

$$\overrightarrow{profile}_{T+1} = \alpha \overrightarrow{profile}_{T} + [\lambda \overrightarrow{topic}_{T} + (1-\lambda)\overrightarrow{doc}_{T}] \quad (6.2)$$

$$\overrightarrow{profile}_{T+1} = \alpha Decay(\overrightarrow{profile}_{T}) + [\lambda \overrightarrow{topic}_{T} + (1-\lambda)\overrightarrow{doc}_{T}] \quad (6.3)$$

Basically, traditional information needs learning methods rely on collecting users' feedback on items, i.e., Webpages, texts, and products. They do not analyze users' information on topics, the rate of topic changes, and the time effects, which may influence the information needs learning results. Therefore, to learn the users' dynamic information needs, we propose three self-profile adaptation methods that consider the topic variation factor, and the time factor. The methods are called P-Time, P-Topic, and P-Topic & Time. As mentioned above, we consider the effect of the time factor and the user's behavior (documents accessed) to adjust the corresponding task profile with the aid of a task-based topic taxonomy. The P-Time and P-Topic methods are similar to the S_P method, but they consider the effect of the time factor and topic profiles, respectively. The P-Topic & Time method adjusts a worker's task profile based on the documents accessed by the worker and their relevance to the topic taxonomy, as mentioned in the section Phase One: Personal Topic Variation Inspection Process (above). The effect of the time factor is also incorporated into the profile adaptation process. The methods are formulated in Equations 6.2 and 6.3. In Equation 6.2, when the parameter λ is set to 0, it is the baseline method, i.e., the S_P method; otherwise it is the P-Topic method. In Equation 6.3, when the parameter λ is set to 0, it is the P-Time method; otherwise it is the P-Topic & Time method. Note that in each equation α is set to 1 during the experiment. Experiments 2 and 3 evaluate the effectiveness of the method for profile adaptation via the collaborative topic variation inspection process. As mentioned above, a worker's information needs can be learned via a personal profile and a collaborative profile (topic-variation profile). The Coll_Topic Variation method uses the collaborative profile adaptation technique defined in Equation 5.3. It adjusts the task profile by a weighted combination of the personal profile and the collaborative profile derived from similar workers. The Coll_Document method is similar to the Coll_Topic Variation method, except that the documents accessed at time $T'_{u_x} + 1$ by similar workers are used to derive the collaborative profile, as shown in the Equation 5.4. In Experiment 2 we evaluate the parameters selected for two collaborative topic variation inspection methods. The parameters δ_V in Equation 5.3 and δ_D in Equation 5.4 are used to adjust the relative weights of the personal profile and the collaborative profile, respectively, in the Coll_Topic Variation and Coll_Document methods. From the values of the two parameters determined in Experiment 2, we select the best values for use in our application domain. Finally, in Experiment 3 we compare four methods (see Table 2) to demonstrate the effectiveness of the information learning method based on the collaborative topic variation inspection process.

Experiment Results

Experiment One: Effects of Profile Adaptation via Self-Topic Variation Inspection

This experiment compares the performance of four methods, namely, S_P, P-Time, P-Topic, and P-Topic & Time under various numbers of top-*N* support documents. The S_P (primitive profiling) method, which is the baseline, learns a user's current information needs from feedback about the recommended information (i.e., documents), and updates the user model for future information filtering. The method only considers a worker's implicit feedback on documents. In contrast, the P-Time and P-Topic methods consider the time factor and topic profiles, respectively. The P-Topic & Time method is a self-profile adaptation method that adjusts the task profile by considering the document profile, the relevant-topic profiles and the time effect simultaneously. Table 3 shows the performances of the four methods in terms of precision, recall, and the F-measure under various numbers of top-*N* documents. Since we want to determine if the system can learn the user's dynamic information needs effectively via the proposed method, we place more emphasis on the precision metric than the recall metric. For the F-measure metric, we set $\beta = 0.5$, (see Equation 6.1) to show the relative importance of precision and recall in order to achieve a trade-off between the two metrics. In addition, we conduct statistical tests to determine whether the observed differences are statistically significant.

Observation 1: Table 3 shows the precision, recall, and F-measure scores under various numbers of top-*N* documents. Clearly, the P-Topic & Time method outperforms the other three methods in each scenario. Figure 6 shows the average precision scores for the four methods. Again, the baseline S_P method yields the least effective performance under various numbers of Top-*N* support documents. This result demonstrates that considering topic profiles and the time

TABLE 2. The methods used in each experiment (The experiments were conducted according to the procedure outlined in the Experiments section).

Method	Descriptions	Parameter setting
Experiment One: Incremental learning technique (baseline technique, traditional IF technique)		
S_P method (baseline method)	• The *S_P* (primitive self-profiling) method is similar to the *Ide_Dec_Hi* algorithm for relevance feedback (introduced in Literature Review). • The *S_P* method is the baseline method. It learns a user's current information needs based on his/her implicit feedback on the textual data, and updates the user model for future information filtering.	S_P method with $\alpha = 1$, $\lambda = 0$ in Equation 6.2
P-Time method	• The *P-Time* method is similar to the *S_P* method, but it also considers the time factor.	P-Time method with $\alpha = 1$, $\lambda = 0$ in the Equation 6.3
P-Topic method	• The *P-Topic* method is similar to the *S_P* method, but it also considers the relevant topics factor.	P-Topic method with $\alpha = 1$ and $\lambda = 0.5$ in Equation 6.2
P-Topic&Time method	• The *P-Topic&Time* method is similar to the *P-Time* and *P-Topic* methods. It considers the time and relevant topics factors simultaneously.	P-Topic&Time method with $\alpha = 1$ and $\lambda = 0.5$ in Equation 6.3
Experiment Two: Parameter Selection for profile adaptation via collaborative filtering methods		
Coll_Topic Variation method	• The *Coll_Topic Variation* method is based on variations in similar workers' topic needs during a specific time period. It is used to derive the collaborative profile. • The method uses δ_V to adjust the relative weights of the personal profile and the collaborative topic-variation profile.	This experiment tries to determine the value of parameter δ_V in the *Coll_Topic Variation* method, as shown in Equation 5.3.
Coll_Document method	• The *Coll_Document* method is based on the documents accessed at time $T'_{ux} + 1$, where similar workers' documents accessed at time $T'_{ux} + 1$ are used to derive the collaborative profile. • The method uses δ_D to adjust the relative weights of the personal profile and the collaborative document profile.	This experiment tries to determine the value of parameter δ_D in the *Coll_Document* method, as shown in Equation 5.4.
Experiment Three: Comparisons of the methods		
S_P method P-Topic&Time method Coll_Topic Variation method (with $\delta_V = 0.7$) Coll_Document method (with $\delta_D = 0.5$)	• Comparison of personal profiling and collaborative profiling by the topic variation methods	None

TABLE 3. Comparison of the self-profile adaptation methods.

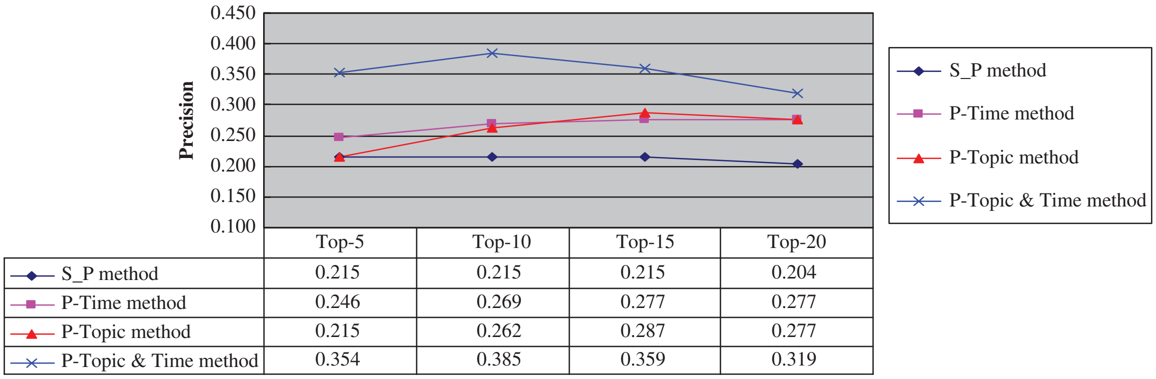
Method	S_P method			P-Time method			P-Topic method			P-Topic & Time method		
Top-*N*	Pre.	Re.	F.	Pre.	Re.	F.	Pre.	Re.	F.	Pre.	Re.	F.
Top-5	0.215	0.030	0.092	0.246	0.035	0.107	0.215	0.022	0.077	0.354	0.041	0.135
Top-10	0.215	0.055	0.129	0.269	0.068	0.160	0.262	0.055	0.146	0.385	0.092	0.225
Top-15	0.215	0.083	0.155	0.277	0.105	0.198	0.287	0.097	0198	0.359	0.132	0.255
Top-20	0.204	0.102	0.161	0.277	0.138	0.219	0.277	0.125	0.214	0.319	0.155	0.252
Average	0.212	0.067	0.134	0.267	0.086	0.171	0.260	0.075	0.159	0.354	0.105	0.217

factor simultaneously during the personal profile adaptation process is effective.

Observation 2: We perform a statistical test to compare learning under the proposed personal profile adaptation methods with that of the baseline S_P method. As shown in Table 4, the results of the P-Time method and P-Topic & Time method are statistically significant at the 0.01 level, i.e., $p < 0.01$, compared to those of the S_P method. However, the results show that the differences in between the F-measure and precision scores of the S_P method and the P-Topic method are not statistically significant, i.e., $t = -1.189$ and $t = -0.463$, respectively. These findings indicate that the time factor is more important than the topic factor in learning users' dynamic information needs.

Experiment Two: Parameter Selection for Collaborative Topic Variation Inspection

Parameter selection for the collaborative topic variation method. The purpose of this experiment is to determine the value of the parameter δ_V in the Coll_Topic Variation method, where δ_V is used to adjust the relative weights of the personal profile and the collaborative topic-variation profile. Note that when δ_V is set to 1, the Coll_Topic Variation method



	Top-5	Top-10	Top-15	Top-20
◆ S_P method	0.215	0.215	0.215	0.204
■ P-Time method	0.246	0.269	0.277	0.277
▲ P-Topic method	0.215	0.262	0.287	0.277
× P-Topic & Time method	0.354	0.385	0.359	0.319

FIG. 6. Comparison of the self profile adapation methods under various numbers of top-*N* document support.

TABLE 4. Statistical test applied to the profile adaptation methods (Compared to the baseline S_P method).

	Method	*Precision* *Significant*	*F-measure* ($\beta = 0.5$) *Significant*
Personal Profile Adaptation Technique	P-Time method	yes, $t = -3.800$**	yes, $t = -4.014$**
	P-Topic method	no, $t = -1.189$	no, $t = -0.463$
	P-Topic &Time method	yes, $t = -3.297$**	yes, $t = -3.719$**

Note. **Significant at $p < 0.01$.

TABLE 5. Effectiveness of the Coll_Topic Variation method under various δ_V values.

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Top-5	0.046	0.185	0.2	0.215	0.338	0.338	0.354	0.369	0.369	0.338	0.354
Top-10	0.062	0.138	0.192	0.292	0.323	0.362	0.385	0.408	0.392	0.385	0.385
Top-15	0.056	0.138	0.215	0.262	0.308	0.354	0.369	0.385	0.359	0.359	0.359
Top-20	0.065	0.146	0.208	0.258	0.304	0.362	0.354	0.358	0.327	0.323	0.319
Average	0.057	0.152	0.204	0.257	0.318	0.354	0.365	0.380	0.362	0.351	0.354

is equivalent to the P-Topic & Time method, which only considers the personal profile. However, when δ_V is set to 0, it is equivalent to the collaborative topic-variation profile. In this experiment, the value of δ_V is systematically adjusted in increments of 0.1. The precision metric is used to evaluate the effectiveness of the methods, and the optimal parameter values with the best results (the highest precision values) are used as the parameter settings of the proposed equations. Table 5 shows the precision rates of the Coll_Topic Variation method with different δ_V values under various numbers of top-*N* support documents.

Observation 3: The results in Table 5 show that the Coll_Topic Variation method with $\delta_V = 0.7$ achieves the best average precision rate under Top-5, Top-10, Top-15, and Top-20 document support. The precision rate of the method increases dramatically from $\delta_V = 0$ to 0.7 and decreases slightly from $\delta_V = 0.7$ to 1, as shown in Figure 7. The experiment results show that, on average, the personal profile is more important than the collaborative topic-variation profile in the Coll_Topic Variation method. However, there are some cases where the performance is better with $\delta = 0.4$ or $\delta = 0.5$; that is, in the given equation, the collaborative profile part is more important than the personal adapted profile part ($\delta = 1$). We discuss these cases further below.

Parameter selection for the collaborative document method. In this experiment we determine the value of the parameter δ_D in the Coll_Document method, where δ_D is used to adjust the relative weights of the personal profile and the collaborative document profile. When δ_D is set to 1, the Coll_Document method is equivalent to the P-Topic & Time method, which only considers the personal adapted profile. However, when δ_D is set to 0, it is equivalent to the collaborative document profile. In the experiment, the value of δ_D is also systematically adjusted in increments of 0.1, and the precision metric is used to evaluate the effectiveness of the methods. The optimal parameter values with the best results (the highest precision values) are used as the parameter settings of the proposed equations. Table 6 shows the precision rates

DOI: 10.1002/asi

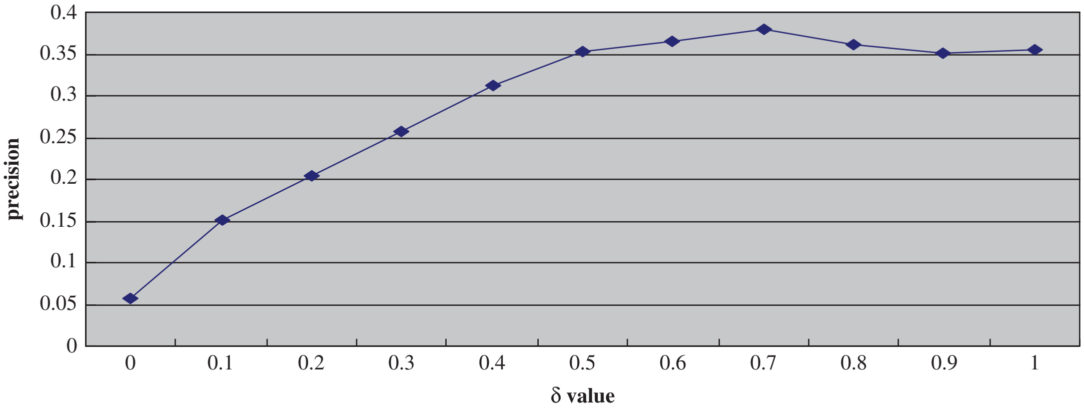


FIG. 7. Knowledge support for Coll_Topic Variation under various values.

TABLE 6. Effectiveness of the *Coll_Document* method under various δ_D values.

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Top-5	0.138	0.231	0.292	0.323	0.323	0.338	0.338	0.338	0.354	0.354	0.354
Top-10	0.185	0.262	0.285	0.346	0.369	0.385	0.385	0.392	0.385	0.385	0.385
Top-15	0.169	0.231	0.303	0.354	0.379	0.369	0.354	0.354	0.359	0.359	0.359
Top-20	0.162	0.25	0.308	0.331	0.346	0.331	0.319	0.319	0.315	0.315	0.319
Average	0.163	0.243	0.297	0.338	0.354	0.356	0.349	0.351	0.353	0.353	0.354

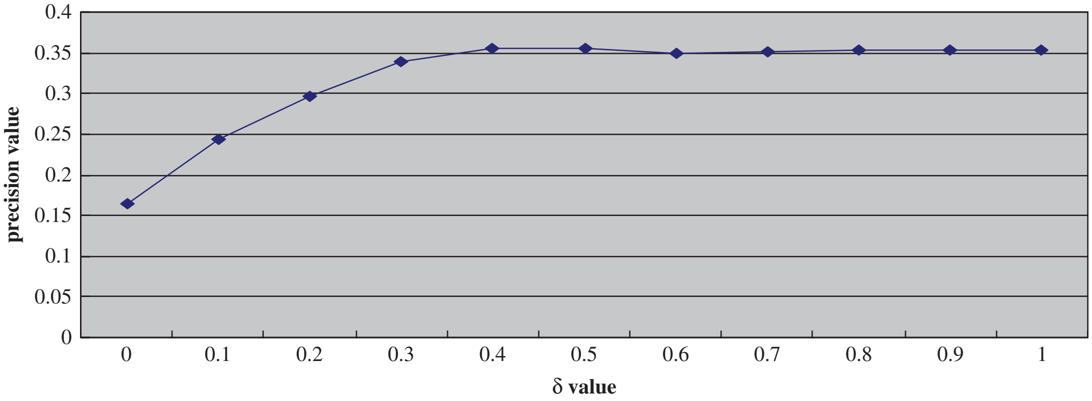


FIG. 8. Knowledge support for Coll_Document under various values.

of the Coll_Document method with different δ_D values under various numbers of top-N support documents.

Observation 4: Table 6 shows that the Coll_Document method with $\delta_D = 0.5$ achieves the best performance (i.e., precision rate). Interestingly, when we set $\delta_D = 0.5$, $\delta_D = 0.8$, $\delta_D = 0.9$, or $\delta_D = 1$, they all yielded similar results. Thus, the curve of the Coll_Document method shown in Figure 8 is smooth and steady from 0.4 to 1. The result indicates that, overall, the collaborative profile of Coll_Document method did not have a significant effect in this experiment.

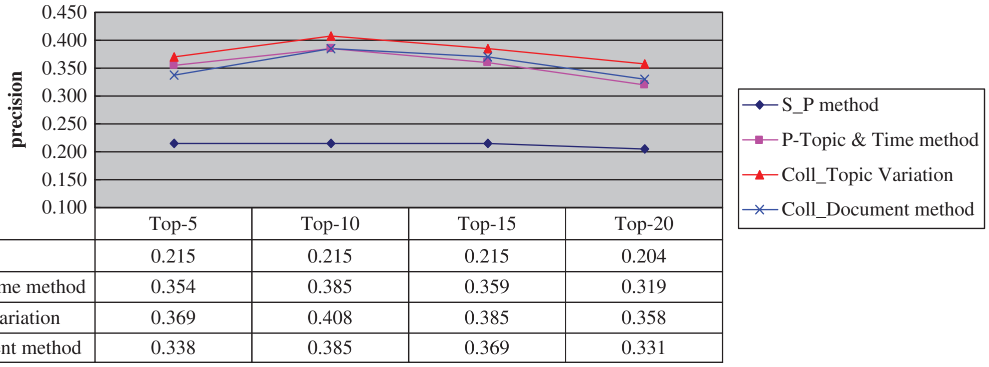
Experiment Three: Comparison of the Two Collaborative Inspection Methods

This experiment compares the performance of task-relevant document support provided by the four methods, S_P, P-Topic & Time, Coll_Topic Variation (with $\delta_V = 0.7$), and Coll_Document (with $\delta_D = 0.5$) methods, under various numbers of top-N support documents. The S_P method is the baseline method, and the P-Topic & Time method is the proposed self-profile adaptation method discussed in Experiment Procedure and Method (above). Neither method considers the effect of the collaborative filtering technique. However, the Coll_Topic Variation method and the Coll_Document method do consider the effect of collaborative profiles generated by workers with similar task-needs. The parameters δ_V and δ_D are used to adjust the relative importance of the worker's personal profile and the collaborative profile, respectively, in the Coll_Topic Variation and Coll_Document methods. The results of Experiment 2 show that $\delta_V = 0.7$ and $\delta_D = 0.5$ yield the best performance. Thus, we adopt the values in Equations 5.3 and 5.4.

DOI: 10.1002/asi

TABLE 7. Comparison of the profile adaptation methods.

Method	S_P method			P-Topic&Time method			Coll_Topic Variation method			Coll_Document method		
Top-*N*	Pre.	Re.	F.	Pre.	Re.	F.	Pre.	Re.	F.	Pre.	Re.	F.
Top-5	0.215	0.030	0.092	0.354	0.041	0.135	0.369	0.045	0.145	0.338	0.040	0.130
Top-10	0.215	0.055	0.129	0.385	0.092	0.225	0.408	0.098	0.237	0.385	0.091	0.222
Top-15	0.215	0.083	0.155	0.359	0.132	0.255	0.385	0.144	0.273	0.369	0.135	0.261
Top-20	0.204	0.102	0.161	0.319	0.155	0.252	0.358	0.181	0.285	0.331	0.160	0.260
Average	0.212	0.067	0.134	0.354	0.105	0.217	0.380	0.117	0.235	0.356	0.107	0.218



	Top-5	Top-10	Top-15	Top-20
S_P method	0.215	0.215	0.215	0.204
P-Topic & Time method	0.354	0.385	0.359	0.319
Coll_Topic Variation	0.369	0.408	0.385	0.358
Coll_Document method	0.338	0.385	0.369	0.331

FIG. 9. Comparison of four methods under various numbers of top-*N* document support.

TABLE 8. Statistical test applied to the collaborative profile adaptation methods (Compared to the baseline S_P method).

	Method	*Precision* *Significant*	*F-measure* ($\beta = 0.5$) *Significant*
Collaborative Profile Adaptation Technique	Coll_Topic Variation	yes, $t = -3.786$**	yes, $t = -4.117$**
	Coll_Document	yes, $t = -3.314$**	yes, $t = -3.783$**

Note. **Significant at $p < 0.01$.

Observation 5: Table 7 shows the precision, recall and F-measure scores under various numbers of top-*N* documents for the different methods. The results show that the Coll_Topic Variation method performs slightly better than the P-Topic & Time and Coll_Document methods under various numbers of top-*N* support documents. Figure 9 shows that the average precision rates of the P-Topic & Time, the Coll_Topic Variation, and the Coll_Document methods are far better than those of the baseline S_P method for various top-*N* retrieval tasks. The results indicate that incorporating a collaborative factor (i.e., profile adaptation based on the topic variations of similar workers) into the profile adaptation process improves the document retrieval performance slightly. Moreover, as shown in Table 8, the results of the proposed collaborative profile adaptation methods, i.e., Coll_Topic Variation and Coll_Document, are statistically significant at the 0.01 level, i.e., $p < 0.01$, compared to those of the S_P method.

Case Investigation

In this work, each evaluation subject involved in executing a task is regarded as a "case." The experiment results show that some of the investigated cases achieved a significant improvement in retrieval performance under specific conditions.

Observation 1: Figure 10 shows three cases (cases 1, 2, and 3) with large variations in topic-needs and a normal case (case 4) with small variations in topic-needs. A higher variation in topic-needs indicates a larger number of changes in information needs during a task's execution. By connecting cases 1, 2, and 3 to the cognitive status of the workers' tasks, we find that the workers are in the topic selection phase, i.e., the pre-focus stage. Hence, their information needs on topics will change frequently. On the other hand, the worker in case 4 is in the task closure phase, so that person has specific information needs for the assigned task. In other words, small variations in topic-needs indicate that the worker's information needs are stable.

Observation 2: Figure 10 also shows the precision rates of the four cases using the Coll_Topic Variation and Coll_Document methods with various δ values. For the Coll_Topic Variation method, cases 1, 2, and 3 perform better under $\delta = 0.4$ or $\delta = 0.5$; that is, the collaborative profile is more important than the personal profile ($\delta = 1$) in the given equation. The normal case (case 4) achieves the

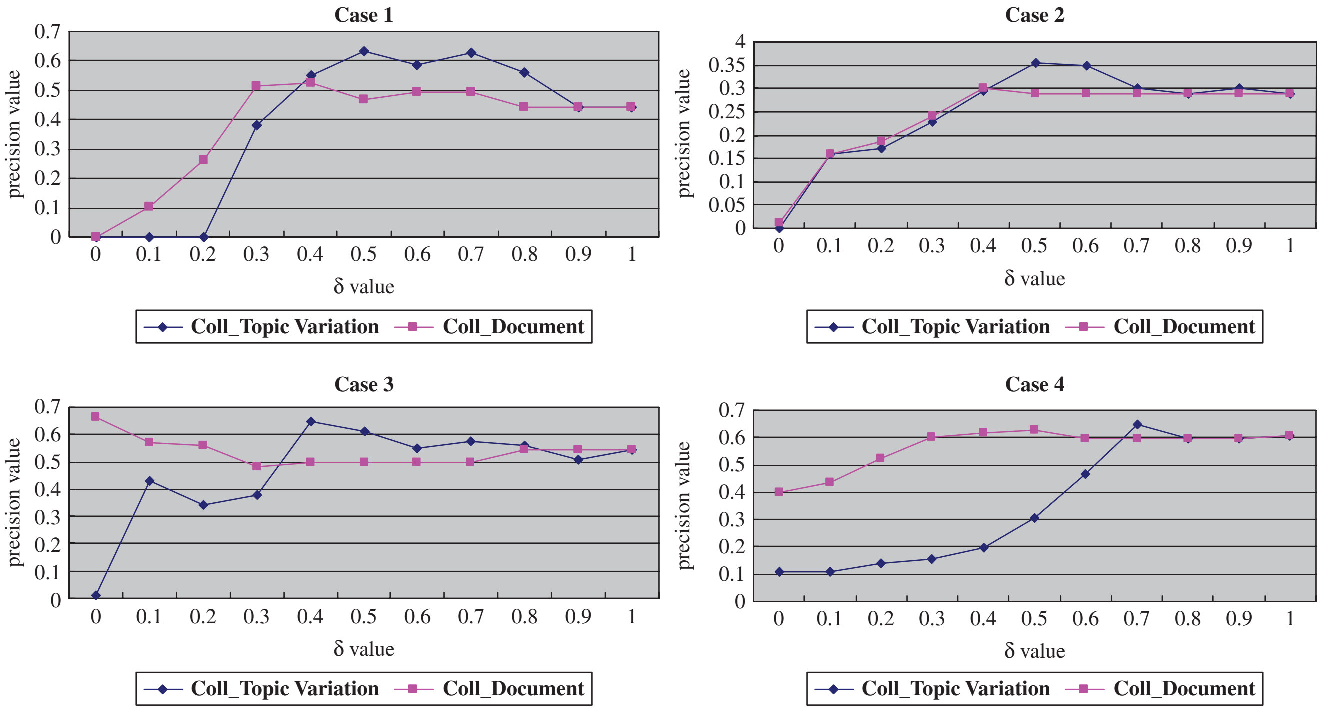
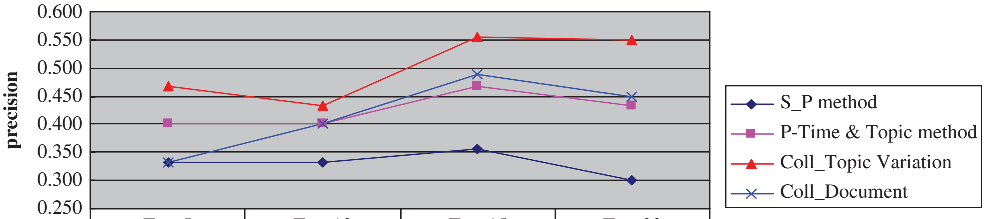


FIG. 10. The four investigated cases.



	Top-5	Top-10	Top-15	Top-20
S_P method	0.333	0.333	0.356	0.300
P-Time & Topic method	0.400	0.400	0.467	0.433
Coll_Topic Variation	0.467	0.433	0.556	0.550
Coll_Document	0.333	0.400	0.489	0.450

FIG. 11. Retrieval performance trends of cases 1, 2, and 3 with various numbers top-*N* support documents.

best performance under $\delta = 0.7$ for the Coll_Topic Variation method, and under $\delta = 0$–0.7 for the Coll_Document method. The results show that it is more effective to give more weight to the collaborative profile during the profile adaptation process in cases with a high degree of variation in topic-needs, i.e., the topic selection phase in the early stage of a task's execution. In contrast, it is more effective to give more weight to the personal profile in cases with small variations in topic-needs, i.e., case 4. In a future work we will include more samples so that we can analyze the effects of different degrees of topic variation.

Observation 3: Figure 11 shows the average retrieval effectiveness for cases 1, 2, and 3 under various numbers of top-*N* support documents. The curve shows that the Coll_Topic Variation method performs better than the S_P, P-Topic & Time, and Coll_Document methods under various numbers of top-*N* support documents. In other words, the Coll_Topic Variation method, which considers collaborative adaptation based on topic-needs variation, is more effective than the other three methods, especially for cases that have large variations in topic needs. Thus, the proposed method is more effective for workers whose information needs change a great deal during a task's execution.

Discussion and Implications

The experiment results demonstrate that the proposed information needs learning method based on personal or collaborative topic variation inspection is effective. First, Experiment 1 shows that the proposed personal profiling methods, i.e., P-Time, P-Topic, and P-Topic&Time, perform better than the baseline S_P method, which is adapted from the Ide_Dec_Hi algorithm used in the traditional relevance feedback (RF) technique. The results confirm that recently accessed documents are more important than older documents in reflecting a worker's current task needs. The precision and F-measure scores of the P-Time method are statistically significant compared to those of the baseline method. However, the precision and F-measure scores of the P-Topic method are not statistically significant compared to those of the baseline method, as shown in Table 4. Thus, the time factor is more important than the topic factor in the profile adaptation process. Furthermore, Experiment 1 shows that the P-Topic & Time method is more effective than the P-Time method. This result demonstrates that considering the topic profiles and the time factor simultaneously during the profile adaptation process is effective. In Experiment 3 we analyzed the variations in workers' task needs on the topic taxonomy to identify workers with similar variations in topic needs over time. Such variations are used to predict future variations in the target worker's topic needs, and to adjust the worker's task profile. We propose two information needs learning methods based on a collaborative topic variation process, namely, the Coll_Topic Variation method and the Coll_Document method. Generally, the two methods yield similar results. That is, workers with similar variations in topic-needs or document interests can be used to predict the target worker's future information needs and improve the retrieval effectiveness. In the cases discussed in above we found that, for workers with a high degree of variation in topic needs, collaborative adaptation of the task profile based on variations in similar workers' topic-needs improved the retrieval performance significantly. Hence, we conclude that the collaborative-profiling factor is more important than the personal profiling factor in cases that involve several changes in topics during a task's performance. The workers are in the topic selection phase, i.e., the pre-focus stage; therefore, their information needs on topics will change frequently, as explained in above. The results indicate that the cognitive status of the workers is also an important factor that influences the system's information needs learning capability. The findings provide us with hints for the design of an effective collaborative information filtering and retrieval model based on workers' problem stages and search behavior patterns.

Nowadays there are increasing demands for more effective enterprise content (document) management (ECM) systems that can go beyond the basic document management functions of KMSs (Paivarinta & Munkvold, 2005; Smith & McKeen, 2003). An effective search solution requires more than a basic search function. The search applications must enhance the system's retrieval capability by including additional techniques. In user-oriented IR research, there is a growing trend to apply information seeking and retrieval techniques in social and organizational contexts to facilitate effective ECM (Ingwersen & Järvelin, 2005; Tyrvinen, Päivärinta, Salminen, & Livari, 2006). This fact motivated us to develop a novel ECM application that considers the user's search pattern and the project's contents to design the search function. As mentioned in the Introduction, effective KM practices require an understanding of workers' information needs when they perform tasks. Most studies propose push-based strategies to provide task-relevant information or codified knowledge items based on the work process without considering users' active search behavior patterns. To address this problem, we try to improve the knowledge retrieval functions in KMSs by incorporating a time factor, i.e., a decay function, into the topic variation inspection process to identify the user's emerging or declining interest in work-task topics at different times. The categories of knowledge in the target domain are also taken into consideration by constructing a task-based topic taxonomy to classify the tasks (topics) of the research domain. Since the taxonomy is based on the tasks in the research domain, so users' information needs can be expressed in terms of the topics instead of keywords. Furthermore, a new task often has some degree of similarity with previous tasks performed in the organization. Therefore, workers who executed similar work-tasks previously may have had similar search patterns for topics. In other words, a worker's future information needs can be predicted by identifying changes in the topics selected by similar workers. Our findings have implications for the design of an effective collaborative information filtering and retrieval model for managing knowledge in enterprises' search systems.

We should acknowledge the limitations of this study. It took 2–3 years to collect and analyze the data, i.e., users' feedback behavior patterns in the presented system, and evaluate our proposed methods. We carried out exploratory longitudinal research in a real-world setting, i.e., a laboratory, where multiple projects were executed by different workers. Therefore, it was not easy to collect data and sample proper cases in the short term. Furthermore, we developed filtering rules based on the characteristics of the research domain, and only a few tasks and subjects were chosen as test subjects in the study. For example, we selected subjects based on the problem stages they were in while executing long-term tasks. That is, each subject was in a different problem stage (i.e., cognitive state) that was identified according to the criteria and the technique used in our previous study. Consequently, the size of the dataset and the number of participants were restricted in the experiments. Second, with regard to the selected tasks (i.e., evaluation cases), we stress that the tasks should be within the same research domain. They cannot relate to topics that differ from those in the task-based domain ontology because we focus on reusing knowledge about previous tasks (i.e., previously executed tasks) to support the execution of a new task. In the future, we will adjust and refine the task-based domain ontology incrementally to include different kinds of topics. We will also demonstrate that the approach can be generalized to different research domains to support workers' information seeking and retrieval activities when executing knowledge-intensive tasks.

Conclusion and Future Work

To put a worker's dynamic information needs into perspective, we propose a topic variation inspection model to facilitate the application of an implicit relevance feedback (IRF) algorithm and collaborative filtering in user modeling. Specifically, we propose a collaborative topic variation inspection approach to enhance the knowledge retrieval function of KMSs by considering a worker's personal search behavior patterns and collaborative search patterns. The results of the experiment demonstrate that the traditional information needs learning method can be improved significantly by analyzing variations in a worker's task-needs for topics (i.e., self-topic needs) over time, as well as by analyzing the collective (i.e., collaborative actors') topic variation patterns and adjusting the worker's profile accordingly. Moreover, the time factor, which is a decay function incorporated into the topic variation inspection process, enhances the system's retrieval performance. This work contributes to the design of knowledge retrieval functions in KMSs and project management systems. The functions are crucial for reusing an organization's knowledge assets efficiently.

White and Jose (2004) analyzed the level of topic change to improve users' search effectiveness. Their study provides us with a good starting point to refine our approach by employing a different topic similarity measure and analyzing the level of topic change to improve the effectiveness of the retrieval functions. In a previous work (Wu, Liu, & Chang, 2008), we proposed a task-stage identification technique for learning a worker's task-needs based on their task stage in order to deliver task-relevant documents effectively. The pilot study showed that a worker's information needs have a low or negative correlation with search sessions/transactions in the pre-focus task-stage, whereas there is at least a moderate correlation with search sessions/transactions in the post-focus stage. The results of the experiment reported in this work also demonstrate that it is more effective to give a higher weight to the collaborative profile during the profile adaptation process for cases with a high degree of variation in topic-needs, i.e., the early stage of a task's execution. In a future work we will consider both topic variation and task-stage issues to enhance the information learning model, as well as the relationships between workers' information needs and their cognitive states. This work only classifies a worker's problem stages into early and late stages based on our previous studies. Therefore, in the future we will perform in-depth analysis to determine how cognitive states and social factors affect users' perceptions of relevance (Bruce, 1994; Campbell & Van Rijsbergen, 1996; Tang & Solomon, 1998; White, 2004). Finally, we will conduct online evaluations to explore the issues related to interactive information seeking behavior or exploratory search activities in terms of the proposed concepts.

Acknowledgments

This research was supported by the National Science Council of Taiwan under Grant NSC 97-2410H-030-030-MY2. We thank the editor-in-chief and anonymous reviewers for their insightful and constructive comments, which helped us improve the quality of the paper.

References

Abecker, A., Bernardi, A., Maus, H., Sintek, M., & Wenzel, C. (2000). Information supply for business processes: Coupling workflow with document analysis and information retrieval. Knowledge Based Systems, 13(1), 271–284.

Agostini, A., Albolino, S., De Michelis, G., De Paoli, F., & Dondi, R. (2003). Stimulating knowledge discovery and sharing. In Proceedings of the International ACM Conference on Supporting Group Work (ACM GROUP'03) (pp. 248–257). New York: ACM Press.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. New York: ACM Press.

Balabanovic, M., & Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. Communications of the ACM, 40(3), 66–72.

Belkin, N.J., & Croft, W.B. (1992). Information filtering and information retrieval: Two sides of the same coin. Communications of the ACM, 35(12), 29–38.

Belkin, N.J., Oddy, R.N., & Brooks, H. (1982). ASK for information retrieval: Part 1. Journal of Documentation, 38(2), 61–71.

Bruce, H.W. (1994). A cognitive view of the situational dynamism of user-centered relevance estimation. Journal of the American Society for Information Science, 45, 142–148.

Byström, K., & Järvelin K. (1995). Task complexity affects information seeking and use. Information Processing and Management, 31(2), 191–213.

Campbell, I., & Van Rijsbergen, C.J. (1996). The ostensive model of developing information needs. In H. Bruce, R. Fidel, P. Ingwersen, P., & P. Vakkari, P. (Eds.), Emerging frameworks and methods: Proceedings of the Second International Conference on Conceptions of Library and Information Science (COLIS'96) (pp. 251–268). Greenwood Village, Colorado: Libraries Unlimited.

Chen, H., Fan, H., Chau, M., & Zeng, D. (2001). MetaSpider: Meta-searching and categorization on the web. Journal of the American Society for Information Science, 52(13), 1134–1147.

Croft, W.B. (1995). What do people want from information retrieval? D-Lib Magazine, Retrieved August 14, 2009, from http://www.dlib.org/dlib/november95/11croft.html

Davenport, T.H., & Prusak, L. (1998). Working knowledge: How organizations manage what they know. Boston: Harvard Business School Press.

Davies, N.J., Duke, A., & Stonkus, A. (2003). OntoShare: Evolving ontologies in a knowledge sharing system. In N.J. Davies, D. Fensel, & van Harmelen, F. (Eds.), Towards the semantic Web-ontology-based knowledge management (pp. 161–176). London: John Wiley & Sons.

De Bra, P., Houben, G.J., & Dignum, F. (1997). Task-based information filtering: Providing information that is right for the job. In Proceedings of INFWET97. Retrieved August 14, 2009, from http://wwwis.win.tue.nl/infwet97/proceedings/task-based.html

Elovici, Y., Shapira, B., & Kantor, P.B. (2006). A decision theoretic approach to combining information filters: An analytical and empirical evaluation. Journal of the American Society for Information Science and Technology, 57(3), 306–320.

Feldman, S. (2004). The high cost of not finding information. KMWorld, 13(3). Retrieved Retrieved August 14, 2009, from http://www.kmworld.com/articles/readarticle.aspx?articleid=9534

Fenstermacher, K.D. (2002). Process-aware knowledge retrieval. In Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS '02) (pp. 209–217). Hawaii. Washington, DC: IEEE Computer Society.

Godoy, D., & Amandi, A. (2006). Modeling user interests by conceptual clustering. Information Systems, 31(4), 247–265.

Godoy, D., Schiaffino, S., & Amandi, A. (2004). Interface agents personalizing Web-based tasks. Cognitive Systems Research, 5(3), 207–222.

Gray, P.H. (2001). The impact of knowledge repositories on power and control in the workplace. Information Technology & People, 14(4), 368–384.

Harman, D. (1992). Relevance feedback revisited. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92) (pp. 1–10). New York: ACM Press.

Ide, E. (1971). New experiments in relevance feedback. In G. Salton (Ed), The SMART retrieval system: Experiments in automatic document processing (pp. 337–354). Englewood Cliffs, NJ: Prentice Hall.

Ingwersen, P., & Järvelin, K. (2005). The turn: Integration of information seeking and retrieval context. Dordrecht, The Netherlands: Springer.

Jansen, B.J. (2005). Seeking and implementing automated assistance during the search process. Information Process & Management, 41(4), 909–928.

Kankanhalli, A., Tanudidjaja, F., Sutanto, J., & Tan (Bernard) C.Y. (2003). The role of IT in successful knowledge management initiatives. Communications of ACM, 46(9), 69–73.

Kelly, D. (2004). Understanding implicit feedback and document preference: A naturalistic user study. Ph.D. Dissertation, New Brunswick, NJ: Rutgers University.

Kelly, D., & Fu, X. (2007). Eliciting better information need descriptions from users of information systems. Information Processing & Management 43(1), 30–46.

Konstan, J.A., Miller, B.N., Malts, D., Herlocker, J.L., Gordon, L.R., & Riedl, J. (1997). GroupLens: applying collaborative filtering to usenet news. Communications of the ACM, 40(3), 77–87.

Kuhlthau, C. (1993). Seeking meaning: A process approach to library and information services. Norwood, NJ: Ablex Publishing Co.

Kwan, M.M., & Balasubramanian, P. (2003). KnowledgeScope: Managing knowledge in context. Decision Support Systems, 35(4), 467–486.

LaBrie, R.C., & St. Louis, R.D. (2003). Information retrieval form knowledge management systems: Using knowledge hierarchies to overcome keyword limitations. In Proceeding of Ninth Americas Conference on Information Systems (AMCIS '03) (pp. 2552–2563). Atlanta, GA:Association for Information Systems

Liu, D.-R., & Wu, I.-C. (2008). Collaborative relevance assessment for task-based knowledge support. Decision Support Systems, 44(2), 524–543.

Mackay, D.M. (1960). What makes the question? The Listener, 62, 789–790.

Middleton, S.E., Shadbolt, N.R., & Roure, D.C. (2004). Ontological user profiling in recommender systems. ACM Transaction on Information Systems, 22(1), 54–88.

Mostafa, J., Mukhopadhyay, S., Lam, W., & Palakal, M. (1997). A multi-level approach to intelligent information filtering: Model, system and evaluation. ACM Transactions on Information Systems, 15(4), 368–399.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. Organization Science, 5(1), 14–37.

O'Leary, D.E. (1988). Enterprise knowledge management. IEEE Computer, 31(3), 54–61.

Paivarinta, T., & Munkvold, B.E. (2005). Enterprise content management: An integrated perspective on information management, In Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS' 05) (pp. 96–96). Washington, DC: IEEE Computer Society.

Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting Web sites. Machine Learning 27, 313–331.

Pons-Porrata, A., Berlanga-Llavori, R., & Ruiz-Shulcloper, J. (2007). Topic discovery based on text mining technique. Information Processing and Management, 43(3), 752–768.

Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137.

Rocchio, J.J. (1966). Document retrieval systems—optimization and evaluation. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Rocchio, J.J. (1971). Relevance feedback in information retrieval. In: G. Salton (Ed.), The SMART retrieval system: Experiments in automatic document processing (pp. 313–323). Englewood Cliffs, NJ: Prentice Hall.

Ruthven, I. (2001). Abduction, explanation and relevance feedback. Unpublished doctoral dissertation, University of Glasgow, Glasgow, UK.

Ruthven, I., Lalmas, M., & van Rijsbergen, C.J. (2003). Incorporating user search behavior into relevance feedback. Journal of the American Society for Information Science and Technology, 54(6), 529–549.

Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4), 288–297.

Salton, G., & McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill Book Co.

Shapira, B., Shoval, P., & Hanani, U. (1999). Experimentation with an information filtering system that combines cognitive and sociological filtering integrated with user stereotypes. Decision Support Systems, 27, 5–24.

Sieg, A., Mobasher, B., & Burke, R. (2004, July). Inferring user's information context: Integrating user profiles and concept hierarchies. Paper presented at the 2004 Meeting of the International Federation of Classification Societies, Chicago, IL.

Smith, H.A., & McKeen, J.D. (2003). Enterprise content management. Communications of the Association for Information Systems, 11, 438–450.

Spies, M., Clayton, A.J., & Noormohammadian, M. (2005). Knowledge management in a decentralized global financial services provider: A case study with Allianz group. Knowledge Management Research & Practice, 3, 24–36.

Spink, A., Wilson, T., Ellis, D., & Ford, N. (1998). Modeling users' successive search in digital environment. D-Lib Magazine, Retrieved August 14, 2009, from http://www.dlib.org/dlib/april98/04spink.html

Tang, R., & Solomon, P. (1998). Towards an understanding of the dynamics of relevance judgements: An analysis of one person's search behaviour. Information Processing and Management, 43 (2/3), 237–256.

Taylor, R.S. (1968). Question negotiation and information seeking in libraries. College and Research Libraries, 29(3), 178–194.

Tyrvinen, P., Päivärinta, T., Salminen A., & Livari, J. (2006). Characterizing the evolving research on enterprise content management. European Journal of Information Systems, 15, 627–634.

Vakkari, P. (2000), Relevance and contribution information types of searched documents in task performance. In Proceedings of the 23rd Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '00) (pp. 2–9). New York: ACM Press.

Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. Information Processing & Management, 39(3), 445–463.

Van Rijsbergen, C.J. (1979). Information retrieval. London: Butterworths.

White, R.W. (2004). Implicit feedback for interactive information retrieval. Unpublished doctoral dissertation, University of Glasgow, Glasgow, UK.

White, R.W., & Jose, J.M. (2004). A study of topic similarity measures. In Proceedings of the 27th Annual ACM Conference on Research and Development in Information Retrieval. (SIGIR '04) (pp. 520–521). New York: ACM Press.

White, R.W., Jose, J.M., & Ruthven, I. (2003). Adapting to evolving needs: Evaluating a behavior-based search interface. In Proceedings of the 17th Annual Conference on Human Computer Interaction. (HCI '03) (pp. 125–128). Narrabundah, Australia: Computer-Human Interaction Special Interest Group (CHISIG) of Australia. Retrieved August 4, 2009, from http://personal.cis.strath.ac.uk/~ir/papers/hci.pdf

White, R.W., Jose, J.M., & Ruthven, I. (2004). An implicit feedback approach for interactive information retrieval. Information Processing & Management, 42(1), 166–190.

White, R.W., & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM '06) (pp. 297–306). New York: ACM Press.

Widyantoro, D.H., Loerger, T.R., & Yen, J. (2001). Learning user interest dynamics with a three-descriptor representation. Journal of the American Society for Information Science and Technology, 52(3), 212–225.

Widyantoro, D.H., & Yen, J. (2005). Relevant data expansion for learning concept drift from sparsely labeled data. Transactions on Knowledge and Data Engineering, 17(3), 401–412.

DOI: 10.1002/asi

Witten, I.H., Moffat, A., & Bell, T.C. (1999f). Managing gigabytes: Compressing and indexing documents and images. Los Altos, CA: Morgan Kaufmann Publishers.

Wu, I-C., Liu, D.-R., & Chang, P.C. (2008). Toward incorporating a task-stage identification technique into the long-term document support process. Information Processing & Management, 44(5), 1649–1672.

Yang, Y., Yoo, S., Zhang, J., & Kisiel, B. (2005). Robustness of adaptive filtering methods In a cross-benchmark evaluation. In Proceedings of the 28th Annual ACM Conference on Research and Development in Information Retrieval. (SIGIR '05) (pp. 98–105). New York: ACM Press.

Ye, Y., & Fischer, G. (2002). Supporting reuse by delivering task-relevant and personalized information. In Proceedings 24th International Conference on Software Engineering (ICSE '02) (pp. 513–523). New York: ACM Press.

Zhou, D., Ji, X., Zha, H., & Giles, C.L. (2006). Topic evolution and social interactions: How authors effect research. In Proceedings of the 15th ACM Conference on Information and Knowledge Management (CIKM '06) (pp. 248–257). New York: ACM Press.