Proceedings of Workshop on Context Modeling and Reasoning (CoMoRea) at 2nd IEEE International Conference on Pervasive Computing and Communication (PerCom'04), March 14 2004, Orlando, Florida. Applied Research Group, Satyam Computer Services Ltd., Bangalore

Context Aware Retrieval in Web-Based Collaborations

Abstract

Availability of any time any device connectivity enables collaboration among geographically distributed team members pursuing a common goal. Providing goal specific information to the team enhances efficiency of the team and for this to happen, context awareness in information provision is essential. We propose a context aware information retrieval system that retrieves, and presents, information of high utility to a user – taking into account user's implicit need for information, capabilities of the user's access device, and other resource constraints of the user. The information retrieved assists users in achieving the collaboration goals. Typically, a user performs multiple activities to achieve a goal and is usually working towards multiple goals. Our information retrieval system analyzes user's history to derive a set of rules using which perceived user as well as team needs and preferences are computed off-line and applied on-line. This process of combining the current context with analyzed context greatly facilitates the retrieval of the most relevant information for users in the considered collaboration domain.

1. Introduction

Context-aware services, i.e., services which provide customized information acquisitions and adaptations to users based on users' contextual information [4], are attracting a lot of attention. One reason is that, today when we try to locate specific pieces of information on the Web, we have to often contend with a plethora of irrelevant information. Also, in an environment with diverse devices being used, it becomes necessary to present the information in formats tailored to suit a user's device characteristics. Context-aware and situation-aware services together enhance the perceived quality of the delivered information.

Consider a research group consisting of Bob, the manager, Alice, the project leader, and Alex, the research assistant. This group interacts with Professor Paul, at a university in a different city. They work closely as a team, discussing the issues at regular time periods and since the group and the university are located at different places, the discussion is via tele-meetings: teleconferences, videoconferences or through email based messaging, with rare face-to-face meetings. The group and the professor are working on multiple research problems – the context aware information-access problem, targeting a submission for a pervasive computing conference, and semantic database modeling with a database conference as the target conference. Bob and the professor are also involved in several management level discussions as well. Tele-Meetings get scheduled ahead of time and the group assembles in Bob's office to initiate a conference call with the professor. This scenario, typical of research collaborations, provides several contextual situations to explore.

Members of the research team mainly perform 3 different activities to reach their common goal – they read to seek knowledge, they communicate to share knowledge, and reach common understanding, a very important prerequisite to reaching the goal. The team achieves the common end goal by setting and achieving smaller goals (milestones). Communication takes place electronically through email, and through the exchange of documents and through phone-based discussions. This domain is described by different activities {reading, email, brainstorming}. Assisting users in aligning themselves to the team's common goal while performing these activities enhances the team's productivity.

The scenario of research collaboration provides challenging situations where importance of a piece of information is dependant upon several factors such as the potential contribution of the document towards achieving the goal of the team, the current activity performed by one or more members of the team, and the recency of the content in the document. Information's utility depends on the context of the user: While performing literature survey a user may find theoretical papers on a topic interesting, whereas the same user, in the implementation phase, may find papers elaborating upon practical issues related to the topic more interesting.

It is important to remember that members of a team are likely to have their own perceptions about their tasks. Generally, interaction through meetings and discussions helps teams to reconcile between differing perceptions. For teams that are constrained by geographical locations, periodic physical interactions are not always feasible. A Web-based system however can assist in users' aligning themselves to the

common goal by providing information that facilitates recall of the team's context. Utility of such contextual information is further enhanced when presented in an appropriate format. A snippet of information about the previous meeting and the progress thereon, ahead of meeting, serves a user's purpose better than a detailed report. Such contextual data acquisition is enabled, if context is embedded in the information retrieval query. A context aware information retrieval system should encompass different dimensions of context of this domain, to retrieve information relevant to the collaboration. Standard IR and IF techniques could be extended for context aware retrieval [3], by considering different dimensions of context in retrieval process.

In this paper, we describe a context-aware retrieval system for assisting collaborators – currently under deployment to assist in the authors' collaborative research activities. In doing so, we extend IR-like relevance-querying techniques so as to adapt or apply them for context-aware querying.

In our proposed system past context sequences are analyzed and aggregated to derive usage trend and user preferences, in the form of rules. A recommender system retrieves information that is relevant to the team's context using the usage trend and the relevance of this retrieved information is further enhanced to suit the individual context, by presenting the information in a suitable format matching the derived user preferences. This process of combining the current context with analyzed context (AC) greatly facilitates the retrieval of the most relevant document/document type for users in the considered collaboration domain. Further, information derived from such combined context sequence analysis makes team working more efficient, bringing the ability to share value added information amongst the team members with respect to the current goal. The system is currently under deployment, assisting the collaboration among the authors.

2. Contexts and their Aggregation

Context is any information that is used to describe a user's environment [4]. *Current context (CC),* is defined by the current activity and the entities associated with it. We consider an activity as a discrete event, for example, receiving an email or reading a paper. Each context is associated with the following entities: activity being performed, user identity, location of the user, data being accessed, device being used, time, network communication channel, and network condition.

Our approach to context-based retrieval is based on aggregating a past sequence of individual contexts and associating the current user activity to those past contexts to determine the best way to deliver information relevant to the current activity. Due to the continuity of user activities aimed at achieving a certain goal, continuity in contexts could be established and exploited for determining and retrieving more relevant information. This aggregation is achieved making use of the repetitive patterns inherent in the activities performed; for example, consider a sequence of email exchanges amongst members in the team. Analysis of this sequence of emails is likely to reveal the common thread across those emails. Likewise, analyzing consecutive meeting minutes could provide contextual information about the current meeting. Context information derived by recognizing patterns has also been explored by Bauer et al in [1]. Bauer et al [1] propose WordSieve, a system where the context of a user's current interest is derived by analyzing the user's web page access patterns. In our system, analysis is used to derive two pieces of aggregated information:

1) Usage Trends: The current focus of the team will be towards achieving immediate milestones. Analysis of the user history helps to identify these implicit, immediate milestones, which

is very important in information retrieval. Figure 1 illustrates how a team achieves a goal by achieving different milestones.

2) User preferences: User has different preferences during different contexts. Generally, user preferences depend on the device being used and activities being performed.

- **Age of the document:** Age denotes how long the document has been present in the system. A user's need for aged documents depends on the activity being performed. The user might prefer documents that are not very old when referring to an email but might not mind if the document is aged while reading a research paper.
- **Type of document:** Documents are presented to a user in different formats. The user's preference of the type of document depends on device and the activity. For example, the user who is about to attend a meeting prefers summaries compared to complete documents, also a user using small device finds reading a summary easier where as user reading a document in his desktop is likely to prefer the original document for reference over the summary.

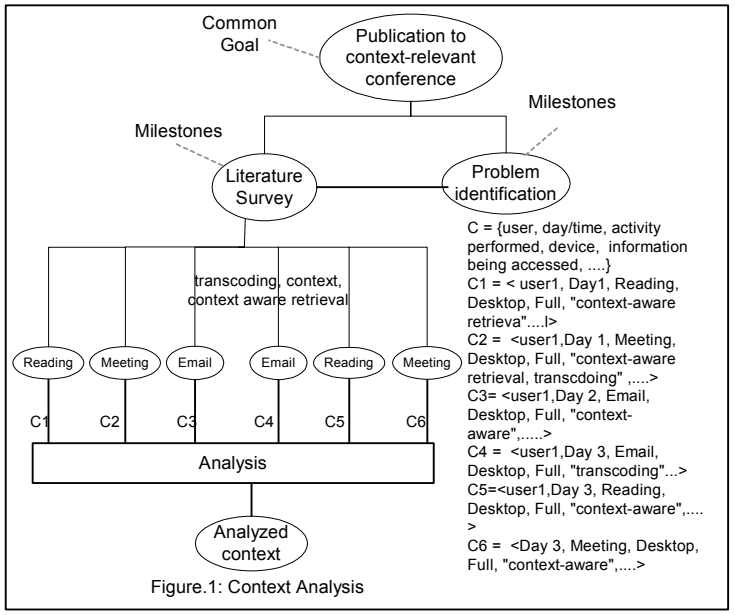

Figure.1: Context Analysis

In summary, by analyzing and aggregating past individual contexts, the *implicit* information and delivery needs of users pursuing a particular activity and the format of presentation of this information in the current context can be captured. When integrated with the information retrieval process, as we discuss in the next section, this technique facilitates the retrieval of context-aware information.

3. Context-aware Information retrieval - approach

The various functionalities of Context Aware Information Retrieval system are depicted in Figure 2, the details of each block are described below:

Rules generated from the Analysis of Past Contexts:

In the current system, we analyze the association between the device used, activity performed and type of document preferred. Such associations, derived by mining past usage patterns are used to derive the user preferences. These preferences are expressed in the form of rules in rule database. For example, consider the context history attributes as shown in Table 1 (only activity, device and type of documents are shown here)

Activity	Device	Type of presented document
Email	Desktop	Full
Email	Desktop	Full
Email	Desktop	Full
Email	Desktop	Summary
Email	Desktop	Summary
Reading	Desktop	Full
Reading	Desktop	Full
Meeting	PDA	Summary
Meeting	Mobile	Voice Summary
Email	Desktop	Summary
Meeting	PDA	Summary

Table 1: Past contexts for rule derivation

The following association rules are identified from the data in Table 1:

Meeting, Mobile => Voice Summary with confidence of 100%

Meeting, PDA => Summary with confidence of 100%

Using similar techniques, user preferences for the age of the documents and rules for usage trend are identified.

Group weight calculation: Documents that are relevant to the team, in addition to their relevance to the context, need to be identified to increase the effectiveness of collaborations. The importance of the document to the team, denoted by group weight (GW), is calculated based on several parameters like hits on the document, the trend of the document usage, which helps in identification of immediate milestones of the team and the roles of the team members who accessed the document. The GW assigned to each document is dependent on the domain specific parameters, and this weight is used in Content-based recommendation step.

Content-based recommendation: In this domain, using a recommender system, that considers the user profiles in addition to the content, increases the precision of the information retrieved. Popescul et al [5] propose a recommender system that uses a three-way aspect model to identify the three-way co-occurrence among users, items and item content. These three entities considered are independent of each other but dependent on a common latent variable. In our research collaboration domain, essentially, the users, documents and the contents of the documents are associated with goal, but are independent of each other. The three-way aspect model is used, with documents, users and contents of the documents being the three different aspects that are connected through goal. The joint probability of co-occurrence of document, user and contents of the document is used to provide recommendations of the relevant documents.

User Preference Matching: The current context of a user is matched with his past aggregated and analyzed contexts (AC) to identify the user's preferences for the current context. This context match is used to further identify the document and its presentable format of the user's particular interest. The user preference rules considered in our system are age preference rule set (AP) and document type preference rule set (TP). Typically, each rule denotes association between different context entities. TP denotes association between user's preference towards a particular type of document when a particular activity is performed using a particular device; and AP denotes association between activity and age of the document.

The measure of context match denotes the confidence of similarity of the past user preferences to the current context. If the current context matches a rule with high confidence, user will be presented with the documents in the format as suggested by the rule. A case where there are no matching rules leads to ambiguity. For example, a CC with <Email, Desktop> results in rule, as derived from Table 1, with a confidence less than the pre-configured threshold. In case of ambiguity, another set of domain-specific rules is applied to identify the user preferences. For example, a different set of rules identified by association between a sub-set of the context entities, in this case the association between device and preferred type of document (TDP) or activity performed and preferred type of document (TAP), is used. If the ambiguity is unresolved even after applying these rules, the system presents users the documents as suggested by the

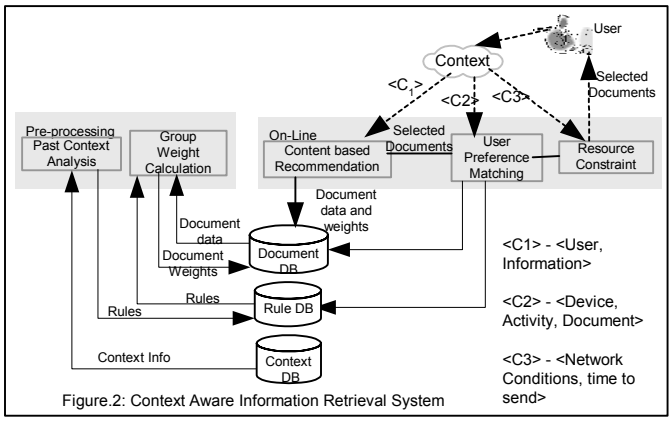

Figure.2: Context Aware Information Retrieval System

default preferences (DP).

Resource Constraints: Delay in information dissemination might render the information irrelevant, if there is a context change. Network conditions may affect the timely delivery of the context data. Time to send is thus an important parameter to be considered for data delivery.
The algorithm used for different components of context aware retrieval is detailed in Appendix A.

4. Evaluation/Early Experiences

Given the current context of a user, the context aware information retrieval system should present the information in the most relevant format. The envisaged system achieves the same by implementing the various modules depicted in Figure.2. Extracting AC from context history and context matching are critical steps in the context aware retrieval process.
In the current implementation, Past Context Analysis and User Preference matching modules have been implemented. A simple query based on keywords has been used to retrieve the documents instead of a recommender system. The current implementation is built around the Tomcat application server using pushlet technology. Servlets that are part of the application server access the relevant information sources to retrieve the information based on the context.
The ongoing collaboration among the authors of this paper gives us a first hand opportunity for subjective evaluation. The documents, the email correspondence, minutes of the meetings, research documents read, involved in the collaboration among the authors were collected. The summaries of the documents were created manually. The context associated with usage of each of the documents was also collected.
Using this data, about 100 different context scenarios were manually constructed. These scenarios were used to train the association-mining engine to generate rules. WEKA (www.cs.waikato.ac.nz/~ml/weka), an open source different data mining tool, was used for mining data to identify association rules. Different context scenarios, involved in our ongoing collaboration were evaluated.
A sample of the training dataset is given in Table 1. Several scenarios with input context limited to < Activity performed, Device used> and output limited to <type of document recommended> and with threshold confidence set to 0.75 is shown in Table 2.

Input	Rules applied	Type of presented document
Meeting, PDA	TP	Summary
Email, Desktop	TDP	Full
Email, Mobile	DP	Summary

Table 2: Context based choices for document presentation

As seen in Table 2, case 2 and case 3 could not be resolved using TP due to ambiguity. In case 3, a default rule, usage of document of the same type that was used in the immediate past context consisting of the same device, was used to resolve the ambiguity.
We are currently extending the scope of the system in several ways: (a) Replacing the default rules by incorporating learning techniques to identify rules to such ambiguities provides enhanced user experience (b) Reducing the latency of access by predicting the user's future context and pre-fetching the relevant documents (c) Providing relevant information from external sources.

5. Conclusion and future work

In this paper, we focused our attention on the problem of developing context-aware information retrieval in web-based synchronous collaborative activities such as tele-meetings. Our overall idea and contributions include the following: (a) defining context to include the information related to activities, device used; (b) identifying user preferences in the domain (c) identifying the immediate focus of the team (d) extending standard recommender system to retrieve relevant information to the team's immediate goal and user content interest (e) presenting the information in the most relevant format aiding in recall while not burdening the device or the underlying network.
Even though we have described various issues related to context aware retrieval information retrieval for synchronous collaboration domain, this approach is applicable to several other domains where regular patterns of activities could be observed.
We plan to continue the ongoing design and end-to-end implementation of the complete system extending the same to (a) extract information from multiple sources including from the web; (b) enrich the context description to include sections of documents of interest to the user; (c) use Artificial Intelligence techniques for learning user preferences to enhance the user experience and (e) consider context prediction and information prefetching.
Literature consists of systems like Dummbo (dynamic ubiquitous mobile meeting board) [2] captures spontaneous, unplanned, and informal activities that take place around a physical whiteboard and provisions for browsing the content on demand by the user. Several systems providing presence-awareness of fellow collaborators are also discussed in the literature. Compared to these systems, the context aware information retrieval system proposed in this paper uses information from such systems to derive the user's implicit need for information and retrieves information assisting the user towards "conceptual awareness" between the collaborators.

References

1. Bauer, T et al., "Exploiting information access pattern for context-based retrieval," *Proceedings of the 2002 International Conference on Intelligent User Interfaces,* ACM Press, 2001.
2. Brotherton, J et al., "Supporting capture and access interfaces for informal and opportunistic meetings,"

Technical Report GITGVU -99-06, GVU Center, Georgia Institute of Technology, January 1999.

3. Brown P.J et al., "Context aware Retrieval: Exploring a New Environment for Information Retrieval and Information Filtering," *Personal and Ubiquitous Computing, **5**,* 4*, pp.253-263, 2001*
4. Dey A.K et al., "The Context Toolkit: Aiding the Development of Context-Aware Applications, " In the *Workshop on Software Engineering for Wearable and Pervasive Computing* , Limerick, Ireland, June 6, 2000
5. Popescul A et al., "Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments," *In Proceedings of the 17'th Conference on Uncertainty in Artificial Intelligence* (UAI 2001), Seatle, WA, August 2001

Appendix A

Context aware information retrieval – Algorithm

In Pre-processing steps, the past contexts are analyzed to arrive at rules and calculate GW. The rules are further used during online process, upon receiving the user's current CC, to retrieve the information most relevant to the user. The algorithm is described as follows:

1. Offline (Pre-processing)

1.1) Calculate Group Weight

Input: Document statistics, context history.

Let the document i^{th} be denoted by d_i, h_i denote the number of hits on d_i and h_{it} denote the number of hits on d_i in time interval T. Let $R = \{r_1, r_2, r_q\}$ be a set of roles in the team, such that the hierarchy is defined in the ascending order.

Group weight for each document is derived as follows:

1. Calculate Hit Factor (HF): HF of the document d_i denotes the popularity of the document in terms of number of hits. HF is given by
   $HFi = h_i / \Sigma_j h_j$
2. Calculate Role factor (RF): Hits by the team member higher in the hierarchy increases the relevance of the document. RF_i is given by $1/r_j$.
3. Calculate Trend Factor (TF): TF of the document d_i denotes the importance of d_i to the immediate focus of the group. TF is given by
   $TF_i = h_{it} / \Sigma_j h_{jt}$
4. Calculate group weight (GWi) of each document by
   $GW_i = (HF_i * TF_i * RF_i)$
5. Update the document database with GW_i

Past Context Analysis

Input: Context history database

1. Mine the history using association-mining technique to identify the association rules set AP, TP, TDP, and TAP.
2. Update the rule database.

2. On-line

Let the current CC be represented by $C_t = \{u, \Delta, l, \tau, a, \omega, \nu, \kappa\}$ where u is the user, Δ is the device, l is the location, τ is the type of information accessed, a is the activity, ω is the information being accessed, ν is the network condition and κ is the time to send. Let $T = \{\tau_1, \tau_2, \tau_l\}$ be a set of type of documents. Let $A = \{a_1, a_2, \ldots, a_p\}$ be a set of activities performed. Let $\Delta = \{\Delta_1, \Delta_2, \ldots, \Delta_q\}$ be a set of devices. Let $U = \{u_1, u_2, u_n\}$ be a set of users, $D = \{d_1, d_2, d_m\}$ be a set of documents accessed by the users, where $d_i = \{d_{i1}, d_{i2}, \ldots, d_{ij}\}$ be a set of different formats of the same document and $W = \{w_1, w_2, \ldots, w_s\}$ be a set of words contained in those documents. Let $Z = \{z_1, z_2, \ldots, z_k\}$ be a set of goals being pursued..

Upon receiving the current context of the user, the online processes are performed in the following order.

2.1) Content-based recommendation

Input: $\{u, \omega\}$ from C_t group weights from document database.Let $P = \{p_1, p_2, \ldots, p_t\}$ be a set of documents retrieved.

1. The joint distribution of an event where the users, documents and the contents to achieve an implicit goal *z* is given by
   $$\Pr(u,d,w) = \Sigma_z \Pr(z)\Pr(u|z)\Pr(d|z)\Pr(w|z).$$
2. The probability of documents that are of interest to n user *u* is given by
   $$\Pr(d/u) \propto \Sigma_w \Pr(d,u,w)$$
3. For each document d_i recommended by the recommender system, calculate the weight by
   $$WT_i = \Pr(d_i/u) * GW_i$$

Output: P, the recommended documents as ordered set, starting from highest weight.

2.2) User Preference matching

Input: $\{\Delta, a\}$ from C_t, document set P from step 2.1 and the rule set TP, AP, TDP, TAP and DP.

Let σ_t be the threshold confidence.

1. Identify context match with AP. Let σ_a be the confidence of the match and α_a be the preferred age given by the rule.
2. If $\sigma_a > \sigma_t$, then output $P' = P - P_a$ where P_a denotes set of documents in P with age > α_a.
3. Identify context match with TP. Let σ_1 be the confidence of the match. Let τ' be the type of document as preferred by the rule.
4. If $\sigma_1 > \sigma_t$, then output P''
5. If $\sigma_1 < \sigma_t$, then identify context match with TDP. Let σ_2 be the confidence of the match. Let τ' be the type of document preferred by the rule.
6. If $\sigma_2 > \sigma_t$, then output P''
7. Repeat 5,6 by replacing TDP with TAP.
8. If $\sigma_2 < \sigma_t$, then output P'' where τ' is derived from DP.

Output: P'', set of documents P' of type τ'

2.3) Resource Constraints

Input: $\{\nu, \kappa\}$ of C_t, P''.

Let the estimated number of bytes that could be sent in time κ, in the current network condition be B.

1. Choose the set of documents in the order of their relevance, such that
   $\Sigma_i S(p_i'') < B$, where $S(p_i'')$ gives the number of bytes in document p_i''.

Output: Set of documents to be presented to the user