



TECHNICAL REPORT 1981
November 2009

SKIPAL Phase 2
Final Technical Report

Bill Deans
Kellie Keifer
Ken Nitz
Laura Tam
SRI International

Michael Carlin
Edward Lai
Doug Lange, Ph.D.
Andrew S. Ling, Ph.D.
SSC Pacific

John Bolton
Bill Graves
Bill Reestman
Northrup Grumman

Approved for public release; distribution is unlimited.

SSC Pacific

TECHNICAL REPORT 1981
November 2009

SKIPAL Phase 2

Final Technical Report

Bill Deans
Kellie Keifer
Ken Nitz
Laura Tam
SRI International

Michael Carlin
Edward Lai
Doug Lange, Ph.D.
Andrew S. Ling, Ph.D.
SSC Pacific

John Bolton
Bill Graves
Bill Reestman
Northrup Grumman

Approved for public release; distribution is unlimited.



SSC Pacific
San Diego, CA 92152-5001

EXECUTIVE SUMMARY

As part of Defense Advanced Research Projects Agency's (DARPA's) Personalized Assistant that Learns (PAL) program, SRI International (SRI) led the Cognitive Assistant that Learns and Organizes (CALO) program to develop learning technologies targeted at the office environment.

In the third year of the program, SRI worked with military personnel to identify PAL technologies that were candidates for transition to military applications. In early 2007, DARPA Director Dr. Tether met with USSTRATCOM Commander GEN Cartwright and identified U.S. Strategic Command's (USSTRATCOM's) Strategic Knowledge Integration Web (SKIWeb) as a candidate for the introduction of technologies from DARPA's Personalized Assistant that Learns (PAL) program. The PAL-enhanced SKIWeb (SKIPAL) work was performed by a tightly integrated team involving SRI, SPAWAR Systems Center (SSC) Pacific, and Northrop Grumman.

CONTENTS

1. INTRODUCTION	1
2. BACKGROUND.....	2
2.1 CALO	2
2.2 CALO-MT	2
2.3 SKIWEB	2
3. SYSTEM ARCHITECTURE.....	4
3.1 DOUBLE HELIX DEVELOPMENT.....	5
3.2 THE SKIPAL WEB SERVICE	6
3.3 PAL TECHNOLOGIES.....	6
3.3.1 Recommendation Engine.....	7
3.3.2 Topic Modeling.....	7
3.3.3 Text-based Classification	9
4. SYSTEM DESIGN AND IMPLEMENTATION	10
4.1 SURVEY PAGE	11
4.2 RECOMMENDATIONS PAGE.....	12
4.3 ENHANCED EVENT PAGE	12
4.4 ALL ACTIVE EVENTS PAGE	18
4.5 DAILY SUMMARY PAGE	19
4.6 QUESTIONS AND ANSWERS PAGE	20
4.7 RECATEGORIZE EVENT PAGE.....	22
4.8 USER PROFILE PAGE.....	23
5. RECOMMENDATION ENGINE EXPERIMENTS	26
5.1 APRIL 2008 EXPERIMENTS (SPIRAL 2.1).....	26
5.2 SPIRAL 2.1 DETAILED EVALUATION	28
5.2.1 Technical Approach	28
5.2.2 Spiral 2.1 Results	30
5.3 AUGUST AND NOVEMBER 2008 EXPERIMENTS (SPIRALS 2.3 AND 2.4)	37
5.3.1 Survey Methodology	37
5.3.2 Analyses.....	38
5.4 SPIRAL 2.3 EVALUATION.....	41
5.4.1 Spiral 2.3 Results	41
6. SUMMARY	49
6.1 CONCLUSIONS.....	49
7. REFERENCES	50

Figures

1. SKIWeb Event page.....	3
2. Existing SKIWeb architecture	4
3. Proposed SKIPAL/SKIWeb architecture	5
4. SKIPAL system configuration for Spiral 2.3 and subsequent spirals	5
5. SKIPAL architecture.....	6
6. SKIPAL component-level design	10
7. SKIPAL Survey page	11
8. SKIPAL Recommended Events page	12
9. SKIPAL Enhanced Event page	13
10. Ask a Question.....	14
11. SKIPAL Answers.....	15
12. Forward a question	16
13. Questions page.....	17
14. View Answers to a Question page	17
15. SKIPAL All Active Events page.....	18
16. SKIPAL Daily Summary page	19
17. SKIPAL Questions and Answers page	20
18. Answer a Question.....	21
19. View Answers to a Question You Asked.....	22
20. Recategorize Event page.....	23
21. SKIPAL User Profile page.....	24
22. SKIPAL Recommends These Events page from April 2008 Experiment.....	26
23. Average ROC by date for the April 2008 Experiment.. ..	27
24. Effect of various parameter values on iLink’s performance as a recommendation engine for eight users on the last day of the April 2008 experiment.....	28
25. Confusion matrix for a given threshold	29
26. ROC curves for individual users in the Spiral 2.1 experiments.	31
27. Precision and recall curves for users in the Spiral 2.1 experiment.	35
28. Sampled precision and recall for the “Max Recommendations” survey data set, representing dates from 27 August 2008 through 21 October 2008.....	39
29. Sampled precision and recall for the “Score Threshold” survey data set, representing dates from 7 November 2008 through 15 January 2009.. ..	40
30. ROC curves for Spiral 2.3 users.	42
31. Precision-recall curves for the Spiral 2.3 (August 2008) experiments.	46

Tables

1. AUC values for Spiral 2.1.....	34
2. AUC values for Spiral 2.3.....	45

1. INTRODUCTION

This is the final report for Phase 2 of the Personalized Assistant that Learns (PAL)-enabled Strategic Knowledge Integration Web (SKIWeb), or SKIPAL, conducted from February 2008 through May 2009. The PAL Military Transition (MT) effort is a multi-pronged effort to transition the technologies developed under the Defense Advanced Research Projects Agency (DARPA) CALO (Cognitive Assistant that Learns and Organizes) program from an office environment to various military software systems.

SKIWeb is an information aggregation system based at U.S. Strategic Command (USSTRATCOM) (Offutt Air Force Base, Bellevue, Nebraska), and available to anyone on Secret Internet Protocol Router Network (SIPRNET). With a user base in the tens of thousands, and a constantly growing number of human and automated contributors, SKIWeb threatens to overwhelm users, causing them to miss the critical information they need amidst the deluge of information not relevant to their tasks.

Search tools are available in SKIWeb to help the user find information. But we hypothesized that PAL technology would be better at learning the kinds of information users are interested in by interpreting their interactions with SKIWeb as implicit or explicit signals. PAL technology could also assist in solving the problem of event identification (i.e., categorization) and expose the relationships between events and SKIWeb users. Both could be leveraged to improve efficiency and quality in USSTRATCOM operations.

This report details the progress on evaluating this hypothesis.

2. BACKGROUND

2.1 CALO

The DARPA-sponsored CALO project brought together leading computer scientists and researchers in artificial intelligence, perception, machine learning, natural language processing, knowledge representation, multi-modal dialog, cyber-awareness, human-computer interaction, and flexible planning. The single research focus of all these experts was to create an integrated system that can “learn in the wild”—that is, adapt to changes in its environment and its user’s goals and tasks without programming assistance or technical intervention.

2.2 CALO-MT

The first 3 years of the DARPA-sponsored PAL program focused almost entirely on cognitive assistance in an office environment. The core projects were the Cognitive Assistant that Learns and Organizes (CALO) and RADAR. Individual projects included the development of learning algorithms, higher level components labeled learning ensembles, and stand-alone learning applications. Many of these components were integrated with Microsoft Office® applications to produce the CALO system and CALO Express (CE). During year 3, concepts for military application of the PAL technologies were formulated and a short movie was produced to stimulate thought by potential end-users.

In August 2007 (PAL year 4), DARPA held its 25th Systems and Technology Symposium, DARPATech. In preparation for DARPATech, a PAL military transition team was formed; operational concepts were refined; and demonstrations were produced for potential Army, Navy, and Air Force applications.

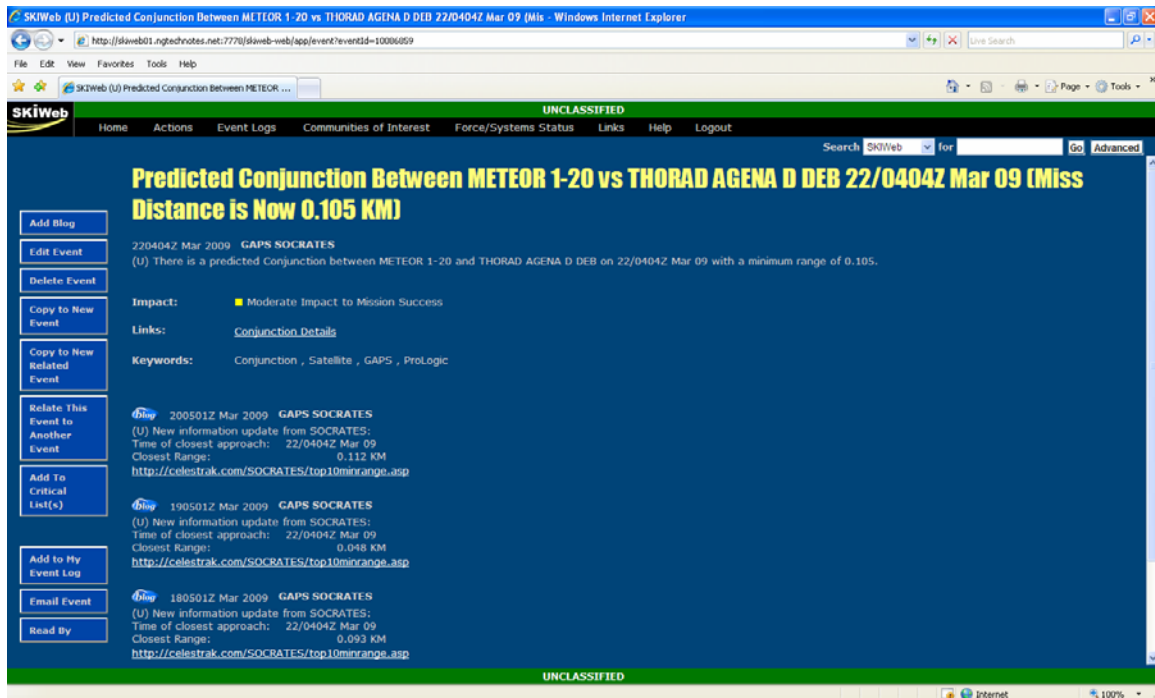
The transition of PAL technologies into SKIWeb is a direct result of the meeting between Dr. Tether, former Director of DARPA, and GEN Cartwright, former commander of USSTRATCOM.

2.3 SKIWEB

The Strategic Knowledge Integration Web (SKIWeb) is a web-based application for integrating and disseminating information on SIPRNET, based at USSTRATCOM (Offutt AFB). Information is contained within an “Event,” a text description that sometimes corresponds to a real-world event. An event could simply encapsulate a news article from open sources, or it could summarize the force readiness of a particular organization. Events have an author, a creation date, and an expiration date. They can also include attachments such as pictures, documents, and links to other websites. During the period between the creation date and expiration date, an event is considered “active.” An example of a SKIWeb event is shown in Figure 1.

Users and automated systems can annotate an event with a “blog.” A blog is a short text description that is appended to the event. Blogs could include corrections, elaborations, questions and answers, or acknowledgements. The event in Figure 1 shows three blogs.

Figure 1. SKIWeb Event page



Depending on the user's role in the organization, he or she will be interested in different subsets of events published on SKIWeb. These events may include news from open or classified sources, force status reports, announcements, or other kinds of information. SKIWeb allows a user to create a dynamic search, called an Event Log, based on words in the events and certain other attributes.

3. SYSTEM ARCHITECTURE

This section describes the architecture of the latest spiral, 2.5. SKIPAL was created as a tightly coupled web application that integrated the data from SKIWeb together with PAL technologies from CALO. The user interface was designed with the same look-and-feel as SKIWeb to provide the users with a familiar environment and to minimize the learning curve. SKIPAL does not duplicate the functionality of SKIWeb. Instead, users are simply forwarded to SKIWeb.

The very first deployed version of SKIPAL was a stand-alone Java[®] application completely separate from SKIWeb. The two applications communicated through a set of web services defined in a shared API (application programming interface). After that, SKIPAL was redeveloped into a web-based application. Since Spiral 2.0, SKIPAL has resided in the same web container as SKIWeb. This tightly coupled nature of SKIPAL with SKIWeb has vastly improved efficiency and reduced the amount of duplicate effort. In particular, it allows SKIPAL to access the SKIWeb database directly without requiring a duplicate SKIWeb database, synchronizing the data with the production database, and maintaining two data access APIs.

Figure 2 shows a simplified view of the existing SKIWeb architecture. The production (PROD) SKIWeb service is a Java 2 Platform, Enterprise Edition (J2EE[®]) web application, residing in a BEA WebLogic[™] container. Persistent data is maintained in the production Oracle[®] database.

Figure 2. Existing SKIWeb architecture

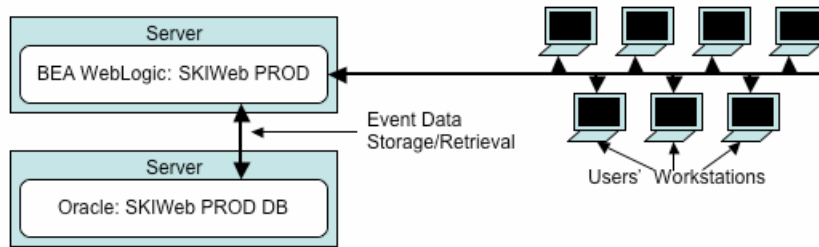
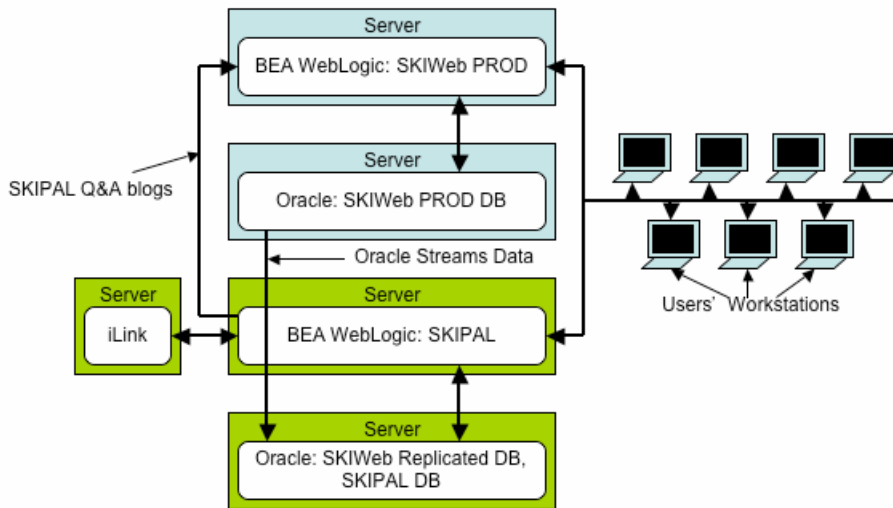


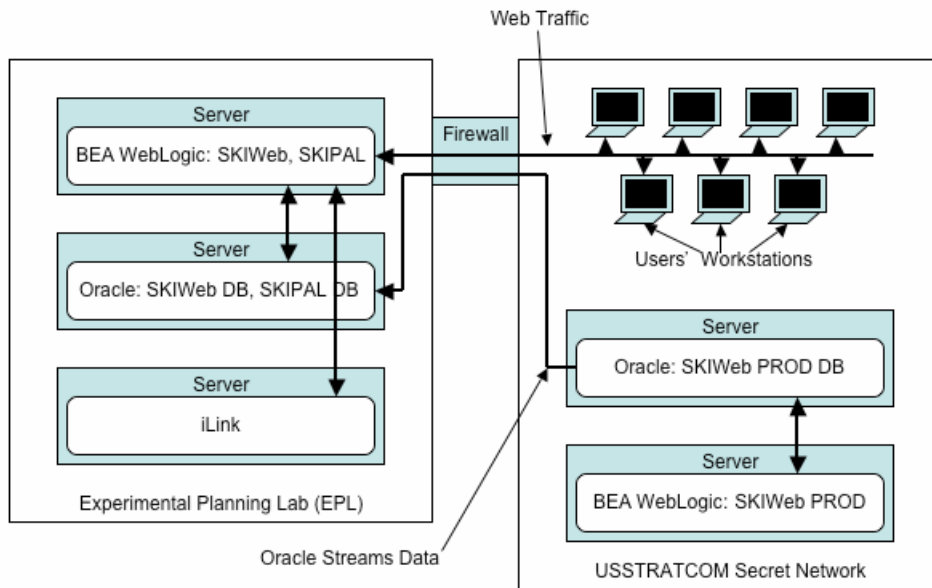
Figure 3 shows the combined SKIPAL/SKIWeb architecture. Database updates are streamed from the production SKIWeb database to a replicated database. SKIPAL interacts with the duplicate SKIWeb database and a database of SKIPAL-specific data residing in the same Oracle instance. In this way, the production SKIWeb is unaffected by SKIPAL and users can choose to interact with SKIWeb, SKIPAL, or both. Using a published web service API, SKIPAL will occasionally blog questions and answers to SKIWeb for the larger community to see. The iLink (see 3.3.2.1) system resides on its own server, and interacts with SKIPAL via its own set of web services.

Figure 3. Proposed SKIPAL/SKIWeb architecture



Since Spiral 2.3, SKIPAL has been operating with the architecture shown in Figure 4. An instance of SKIPAL is isolated in the Experimental Planning Laboratory (EPL) at USSTRATCOM. The addition of a firewall allows a limited number of users to access the SKIPAL server. Since SKIWeb moved to a new version, the SKIPAL team has not been able to get permission from USSTRATCOM information assurance authorities to blog from SKIPAL to the production SKIWeb.

Figure 4. SKIPAL system configuration for Spiral 2.3 and subsequent spirals



3.1 DOUBLE HELIX DEVELOPMENT

SKIPAL exposes USSTRATCOM users to PAL technologies to elicit requirements for transitioning PAL into SKIWeb. The requirements are generated from what DARPA calls a “Double

Helix” event. A “Double Helix” event allows the users to see interim capabilities and provide feedback to the development and management team.

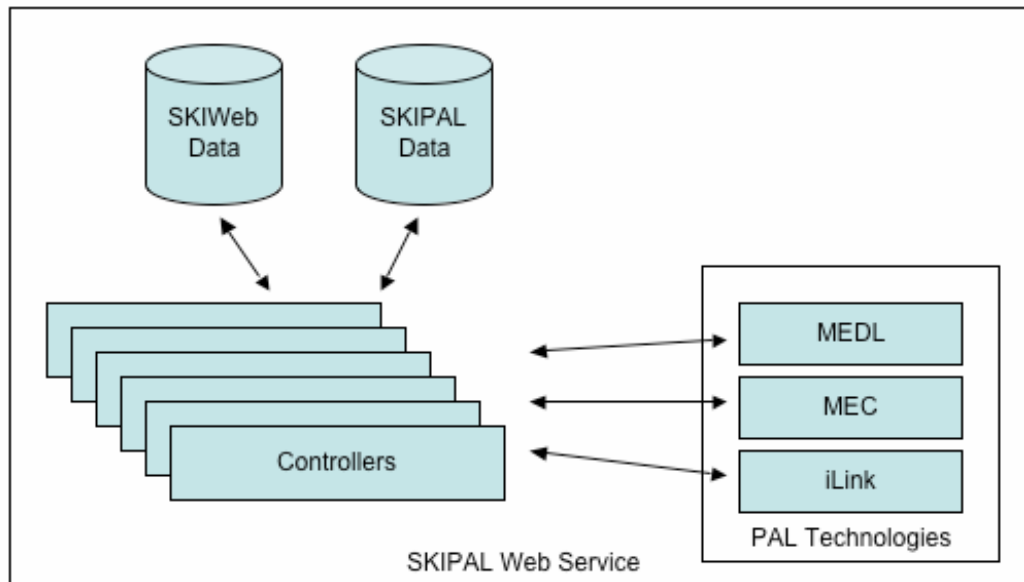
During each Double Helix event, the SKIPAL project team reviewed the current version of the software and encouraged user feedback. Then we presented mockups of suggested changes and new features and invited feedback again. Afterwards, the developers used the feedback to update the software requirements and develop the next iteration of SKIPAL. The Double Helix events have produced invaluable information. As a result of these discussions, several PAL technologies were identified as inappropriate, while others have been made substantially more useful and robust.

A “casualty” of the Double Helix process has been any kind of overarching, formal system architecture. This effort emphasized rapid development and rapid response to user feedback, while controlling complexity and risk. Scalability and fault-tolerance were explicitly not addressed, and we expect that these features will need to be addressed during the acquisition process.

3.2 THE SKIPAL WEB SERVICE

SKIPAL is a standard J2EE web application (“war” file). Its web services are shown in Figure 5. It comprises a number of controller objects that handle the display of web pages and the processing of user actions. The controllers share data with each other through the SKIPAL database, and communicate with shared instances of the PAL components.

Figure 5. SKIPAL architecture



3.3 PAL TECHNOLOGIES

In considering the myriad of PAL technologies that could be integrated into any system, including SKIWeb, it is important to realize that the value of streamlining a user task (through a more efficient interface, automation, or some other means) is bounded by the interest the user has in doing that task in the first place. Of the many CALO products, four were initially identified as being of possible value to the SKIWeb community. Of these four, three turned out to be very appropriate for SKIPAL.

The three appropriate technologies were (1) Recommendation Engine (for learning user interests), (2) Topic Modeling (for identifying relations between people and events, questions and answers), and (3) Text-based Classification (for automatically identifying the categories of an event). Only Task Assistant, a software package that supports creating, managing, automating, and delegating task lists, was identified as inappropriate.

3.3.1 Recommendation Engine

The Recommendation Engine uses feedback from the user to build a mathematical model. The model is applied to a list to reveal items that are of the most interest to the user.

3.3.1.1 MEDL Recommendation Engine

The MEDL (My Event Decision Learner) Recommendation Engine learns the keywords in events the user likes and dislikes and then recommends new events to the user. It silently monitors the events the user reads and accepts positive and negative feedback on an event from the user. It does not factor in the interest profiles of the other users because our tests showed that SKIPAL users do not want to see events based on other users with similar interest profiles.

3.3.1.1.1 Operation

MEDL is a text-based learning system. When MEDL is trained, the words from the event are extracted, common words (“stop words”) are removed, and the remaining words are associated with interest or disinterest.

From the SKIPAL event page, the user has the opportunity to vote “thumbs up” or “thumbs down” on an event. This direct training teaches MEDL whether or not the user likes the event. MEDL also uses an indirect method of learning. When a user reads an event, MEDL learns that the user is interested in this event and to show more events similar to it.

On the User Profile page, the user can add words of interest to MEDL’s training to give more weight to events containing those keywords and recommend them to the user. The user can also specify words of disinterest by prefixing keywords with a minus sign (“-“). MEDL will reduce the weight of events containing words of disinterest and not recommend them to the user.

There are also categories such as countries and regions in the User Profile that the user can select that will tell MEDL that these types of events are important. MEDL will expand the selected country/region designation to include important cities and people in that region to include events that may specify only a city or person and not the country. If necessary, the user can correct MEDL when reading the event.

MEDL uses the training to generate a score between zero and one for each new event, where a score of one means the user would like the event. These scores are displayed on the SKIPAL Recommends page. Clicking on the “Score” column heading will change the ascending/descending sort order of the events.

3.3.2 Topic Modeling

Topic modeling identifies relationships between similar types of information and groups them. It is used in SKIPAL to relate similar events and users who have authored similar events or blogs. Such users could be considered “experts” on the event topic. The similar events and contributors are part of the SKIPAL Decorated Event page. SKIPAL also uses this information to suggest users who might be able to answer a question that is asked on an event.

3.3.2.1 iLink

iLink is a topic modeling system designed to support social networking applications. A user's expertise is defined by iLink as the user's knowledge of or interest in different kinds of documents. iLink tracks the user's interaction with similar kinds of documents. It accepts both explicit signals such as "I am interested in this document" or "I don't like this article" and implicit signals such as authoring, blogging on, reading, or adding an event. The weight of these signals can be adjusted to model different domains or use cases.

iLink has been successfully integrated into several military systems, including Platoon Leader, a web forum for Army platoon leaders to share information. Therefore, it seemed that iLink would also perform well with SKIWeb.

In SKIPAL, iLink has been used in several areas: recommending events (later replaced by MEDL), suggesting similar events and subject matter experts on the Enhanced Event page, and suggesting similar Q&A pairs and subject matter experts on the Ask a Question page. A technical paper on iLink is available (reference [1]).

3.3.2.1.1 Using iLink as a Recommendation Engine

iLink performed poorly as a recommendation engine. This was due in part to some design choices in iLink that conflict with the SKIWeb usage paradigm. SKIWeb is not a social networking site or a web forum in the typical sense because the relevancy of an event to a user is determined by the age and content of the event in SKIWeb. iLink does not factor document age into its model. Therefore, an event that was popular years ago may continue to be recommended (if the event is still active), even though its age has rendered it irrelevant to the SKIPAL user.

iLink also presented some problems in determining expertise in SKIPAL. Similar events read by the user are an important implicit signal in determining the relevance of an event. However, iLink also assumes that a user's expertise in that subject will increase as a result of reading that event. The side-effect is that iLink will consider that user to be a subject matter expert on events of that type in the Enhanced Event Page (see section 4.3). While this would be perfectly natural in a web forum, it does not fit the SKIWeb model. The role of each SKIWeb user is defined by his/her job. Most SKIWeb users are either consumers or producers of specific types of events. For example, users who routinely read about weather-related events typically do not generate those kinds of events in SKIWeb. Their role is to consume the information provided by the subject matter experts. Therefore, our solution to this problem was to exclude reading an event as a signal provided to iLink.

3.3.2.1.2 Suggesting Similar Events and Contributors

iLink is used to recommend people and similar events for the Enhanced Event Page. The related events are identified by their similarity in content. Unlike recommended events, the age of an event is not a significant factor in determining similar events. The list of suggested contributors is based on the content of the event matched to the expertise information iLink stores for each user. As previously mentioned, the interpretation of expertise can vary widely depending on the type of signals received by iLink. Since SKIPAL is suggesting contributors, iLink is only notified when a user creates, updates, or blogs on an event. As a result, the list of contributors that iLink suggested were much more accurate according to user feedback but we have not yet designed or conducted any sort of objective performance evaluation.

3.3.2.1.3 Question and Answer Page

iLink was originally designed for questions and answers. When a question is asked on an event, iLink will suggest people who may be able to answer the question. As users answer questions or create events or blogs on similar topics, their modeled expertise in the topic increases and they become more likely to be recommended for similar questions in the future.

iLink will also recommend related questions that have been answered. Sometimes a similar question will have the answer the user needs. Or perhaps, the question is a duplicate and has already been answered.

The Q&A functionality was deployed in the USSTRATCOM EPL on March 30.

3.3.3 Text-based Classification

The Multi-Class Event Category (MEC) Classifier is a classifier that will classify events into multiple categories based on prior training by users. The categories represent common areas of interest to the SKIPAL community and are the same for each user. As such, every user benefits when MEC is trained. New categories can be added by contacting the SKIPAL administrator. A maximum of about 20 to 25 categories are possible. MEC is a multi-class classifier, which allows an event to be assigned up to three different categories.

3.3.3.1 Training

MEC uses the Maximum Entropy classifier from the Machine Learning for Language Toolkit (MALLET) code base in a hierarchical fashion. First, MEC extracts the text in an event and removes all of the common words. Then, the event's category and its associated words are sent into the "Main" categorizer that uses all the categories at once. Next, a binary classification engine that trains with the training category against all other categories combined. This hierarchical system creates a classification engine for each category and a "Main" classification engine to determine which separate classification engines to run. This allows MEC to keep the number of classification engines to run during classification for each event low.

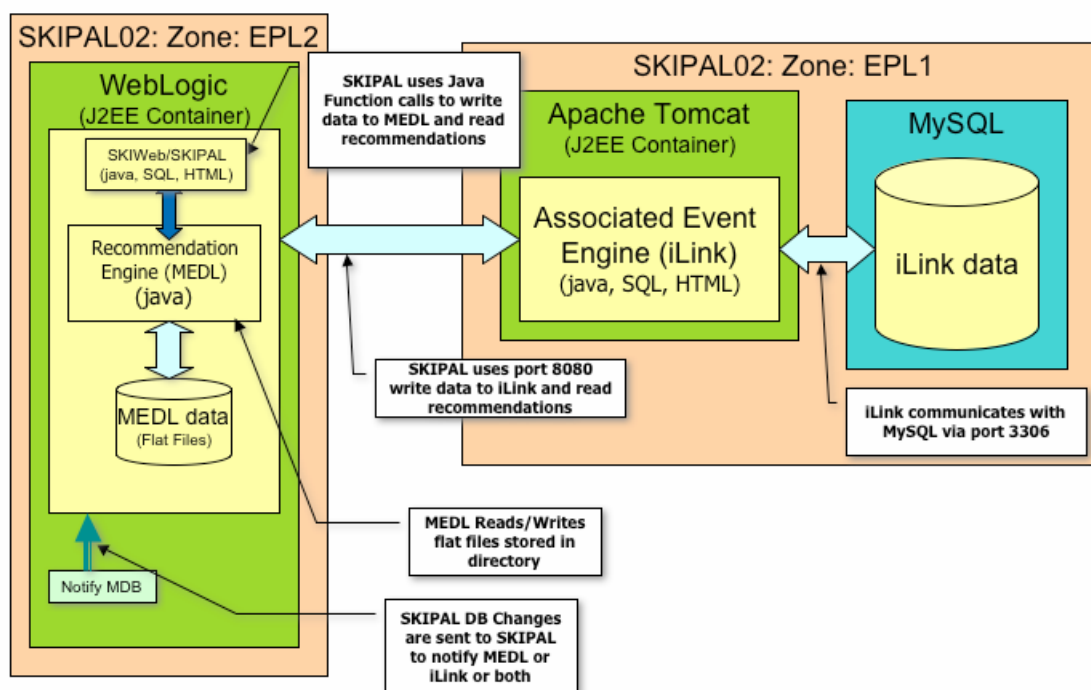
3.3.3.2 Classification

To classify an event, all of the words are extracted from the event and the common words are removed. MEC uses the "Main" classification engine to determine which categories have the highest probability of being correct. Then it calculates a threshold, filters on each category's probability, and sends the remaining categories through their binary classifiers. The top three categories or less are displayed. If MEC cannot determine the category of an event, then it returns "unknown."

4. SYSTEM DESIGN AND IMPLEMENTATION

Figure 6 shows a high-level view of the SKIPAL software components. SKIPAL is composed of Java Enterprise web services, the MEDL recommendation engine, the MEC Classifier, and iLink running alongside an instance of SKIWeb. MEDL stores its persistent data in flat files. iLink has its own web service provided through Apache Tomcat and persisted by a MySQL® database. SKIPAL interacts with iLink via iLink's web services.

Figure 6. SKIPAL component-level design



In Phase 2, SKIPAL incorporated the Spring Web MVC framework to facilitate future integration efforts with SKIWeb. Spring Web MVC is a web framework based on the Model-View-Controller (MVC) design pattern.

New events created in SKIWeb activate a database trigger that captures the id numbers of the new events as a Java message and places it into the Java Messaging Service (JMS) queue for delivery to SKIPAL. SKIPAL uses the event id numbers to query the SKIWeb database for the complete event data and user data as needed. In addition, SKIPAL has its own database to store data on SKIPAL user profiles, questions asked, and accepted answers. SKIPAL also has tables for collecting metrics from event re-categorizations and recommendations relevancy.

The following sections focus on the user interface design. The goals of the user interface design were to create a useful and intuitive PAL enhanced portal to SKIWeb data and to provide mechanisms of eliciting feedback from the user on SKIPAL's performance.

4.1 SURVEY PAGE

After the user logs into SKIPAL, clicks on the Recommendations tab, or clicks on the All Active Events tab, the user is presented with the Survey page (Figure 7).

Figure 7. SKIPAL Survey page

UNCLASSIFIED

SKIPAL Survey

Please take a moment to complete this brief survey. Your response will help SKIPAL tailor its recommendations to your interests.

You will be redirected to the *Recommendations* page upon completion of the survey.

Is the event below relevant to your interests? Yes No

SKIPAL believes the event below belongs to the following Space Ops categories:

Has SKIPAL categorized the event correctly? Yes No

If incorrect, please select all the correct categories:
You may select up to 3 event categories by holding down the **SHIFT** or **CTRL** key.

- Afghanistan
- Africa
- Asia
- Business
- China
- Congo
- Crime
- Disaster
- Economy
- Entertainment

Submit

Predicted Conjunction Between ORBCOMM FM19 vs FENGYUN 1C DEB 10/1244Z Mar 09 (Miss Distance is Now 0.1 KM)

101244Z Mar 2009 GAPS SOCRATES

(U) There is a predicted Conjunction between ORBCOMM FM19 and FENGYUN 1C DEB on 10/1244Z Mar 09 with a minimum range of 0.1.

Impact: ● Moderate Impact to Mission Success

Links: [Conjunction_Details](#)

Keywords: Conjunction , Satellite , GAPS , ProLogic

Expiration: 111244Z Mar 2009

Event Entered: 100501Z Mar 2009

Event Date: 101244Z Mar 2009

UNCLASSIFIED

The Survey page is used to elicit feedback from the user for the purpose of analyzing SKIPAL's performance. Events are randomly chosen from all currently active events. SKIPAL improves on recommending and categorizing events with more user feedback. The user is not allowed to continue to the Recommended Events page or All Active Events page unless the survey is completed and submitted. Users can opt out of the survey at anytime. To do so:

1. On the upper-right corner of the main page, click **Update Profile**.
2. Next to "Do you want to participate in the survey?", select **No**.
3. Click **Submit**.

4.2 RECOMMENDATIONS PAGE

The Recommended Events page (Figure 8) shows a list of events as recommended to a user by SKIPAL.

Figure 8. SKIPAL Recommended Events page

Date	Title	Blog	Question	Categories	Score	
091148Z Mar 2009	(U) Brown says Northern Irish progress will not be halted Participants: GAPS SKIWeb II News Injector		[1 question]	---	0.936	[thumbs-up] [thumbs-down] [red X] [envelope]
111316Z Mar 2009	(U) Predicted Conjunction Between METEOR-3M vs THOR ABLESTAR DEB 11/1316Z Mar 09 (Miss Distance is Now 0.1 KM) Participants: GAPS SOCRATES	[1 blog]		Space Ops	0.933	[thumbs-up] [thumbs-down] [red X] [envelope]

New SKIPAL users will not have any events recommended until the user has read at least five events from the All Active Events tab, participated in the survey, and/or clicked on the **thumbs-up** and/or **thumbs-down** button on a few events. The first two columns in the list of SKIPAL Recommended Events are the event date and the event title. The next two columns, Blog and Question, will display an icon next to an event that has a blog or a question. The Categories column displays up to three categories assigned to the event by the MEC. The Score column displays the probability that the event is relevant to the user as computed by the MEDL recommendation engine. Finally, each row has a set of four buttons on the right-hand side. The **thumbs-up** and **thumbs-down** buttons allow the user to inform the recommendation engine that this event is more or less relevant, respectively. Clicking the **thumbs-down** button also removes the event from the list. The **red X** button filters the event without indicating relevance. Clicking the **envelope** icon allows a user to refer this event to another user. Clicking on the **event title** will open a new window or tab containing the Enhanced Event page for that event (see section 4.3). By default, events on the Recommendations page are sorted in decreasing MEDL score but can be resorted by clicking on any of the other column headings, except the Categories column.

Phase 2 included two major iterations of this page. The first iteration limited the number of recommendations to a user-specified number and is referred to in Section 5 as the Max Recommendations Scheme. Also, it showed only one category and did not display the score.

The second and current iteration allows the user to set a threshold value for the recommendation engine. If an event's score is greater than or equal to the threshold, then it is displayed. Research conducted in search engines has suggested that displaying the score can confuse the user since scores are often unbounded and relative. For example, one search might have an initial result score of 10 while another might return a score of 10 million. Therefore, most search engines today hide the scores. However, SKIPAL scores are normalized to [0, 1], and exposing the score allows both the recommendation engine and the user to influence the number of items recommended.

4.3 ENHANCED EVENT PAGE

The SKIPAL Enhanced Event page (Figure 9) is similar to the SKIWeb event page but includes additional information from the MEC classifier and iLink. The Event Category, Similar Events, and Suggested Contributors sections are enhancements to the standard SKIWeb event page. The Event

Category is determined by the MEC. The Similar Events and the Suggested Contributors content come from iLink. The event's category is shown in the upper right-hand corner. Clicking on the **Recategorize Event** button brings up the page shown in Figure 20. The **Ask Question** button on the lower left of the page is another enhancement to the basic SKIWeb page. Clicking on the **Ask Question** button brings up the Ask a Question page shown in Figure 10. If one or more questions have been asked about this event, then a **View Questions and Answers** button will be displayed below the **Ask Question** button.

Figure 9. SKIPAL Enhanced Event page

UNCLASSIFIED

<div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-bottom: 2px;">Add Blog</div> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-bottom: 2px;">Copy to New Event</div> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-bottom: 2px;">Copy to New Related Event</div> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-bottom: 2px;">Relate This Event to Another Event</div> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-bottom: 2px;">Add To Critical List(s)</div> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-bottom: 2px;">Add to My Event Log</div> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-bottom: 2px;">Email Event</div> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-bottom: 2px;">Read By</div> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-bottom: 2px;">Ask Question</div> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-bottom: 2px;">View Questions and Answers</div> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center;">Help</div>	<div style="background-color: #004a99; color: white; padding: 5px;"> <h2 style="margin: 0;">BBC News News Front Page World Edition : UK to see 'more' swine flu cases</h2> <p style="margin: 0;">301202Z Apr 2009 BBC News News Front Page World Edition</p> <p style="margin: 0;">(U) Britain will see "many, many more cases" of swine flu but most people will recover, a chief government adviser says.</p> <p style="margin: 0;">Impact: ● Little/No Impact to Mission Success</p> <p style="margin: 0;">Links: Click Here for Full Story UK to see 'more' swine flu cases</p> <p style="margin: 0;"> 302153Z Apr 2009 SKIPAL</p> <p style="margin: 0;">(U) Michael Carlin asked the following question: Is it time to start panicing???</p> <p style="margin: 0;">Ken Nitz gave the following answer: yes</p> <p style="margin: 0;">Updated: 302153Z Apr 2009</p> <p style="margin: 0;">Expiration: 012153Z May 2009</p> <p style="margin: 0;">Event Entered: 301202Z Apr 2009</p> <p style="margin: 0;">Event Date: 301202Z Apr 2009</p> </div>	<div style="background-color: #004a99; color: white; padding: 5px;"> <p>Event Category Health</p> <div style="background-color: #004a99; color: white; padding: 2px; text-align: center; margin-top: 5px;">Recategorize Event</div> </div> <div style="background-color: #004a99; color: white; padding: 5px; margin-top: 5px;"> <p>Similar Events 1</p> <ul style="list-style-type: none"> ● 042002Z May 2009 (U) BBC News News Front Page World Edition : Nine more UK cases of swine flu ● 280402Z Apr 2009 (U) BBC News News Front Page World Edition : Swine flu prompts travel warning ● 300402Z Apr 2009 (U) BBC News News Front Page World Edition : Three new swine flu cases in UK ● 291202Z Apr 2009 (U) BBC News News Front Page World Edition : Three new swine flu cases in UK ● 272002Z Apr 2009 (U) BBC News News Front Page World Edition : Swine flu cases confirmed in UK <p style="text-align: center; margin-top: 5px;"> << first < prev 1 2 next > last >> </p> </div> <div style="background-color: #004a99; color: white; padding: 5px; margin-top: 5px;"> <p>Suggested Contributors 1</p> <p>Mr CIV Michael Carlin (COMM (P): 619-553-4681)</p> <p>CIV Ryne Hobbs (DSN (P): 402-555-1212)</p> <p>Hari Seldon (COMM (P): 123-456-7890)</p> <p>Mr KTR Skiweb Tester (COMM (P): 123-4567)</p> <p>Edward Lai (DSN (P): 619-553-5209)</p> <p style="text-align: center; margin-top: 5px;"> << first < prev 1 2 next > last >> </p> </div>
--	---	--

UNCLASSIFIED

Figure 10. Ask a Question

UNCLASSIFIED

Maj Frank Gas's SKIPAL

Ask a Question

about the following event:

301202Z Apr 2009 (U) BBC News | News Front Page | World Edition : UK to see 'more' swine flu cases [Blog](#) [1 blog]

* Question: Are we going to die from the swine flu?

3961 characters left

Submit Question

SKIPAL users can ask questions about an event by clicking on the **Ask Question** button on the Enhanced Event page. When a question is submitted, SKIPAL provides similar questions that have been answered (Figure 11) to the user.

Figure 11. SKIPAL Answers

The screenshot displays the SKIPAL interface with a dark blue background and a green header bar. The header bar contains the text "UNCLASSIFIED". Below the header, the page title "Maj Frank Gas's SKIPAL" is shown in yellow. The main heading "SKIPAL Answers" is also in yellow. Underneath, the section "Your question:" is followed by a text input field containing the question: "* Question: Are we going to die from the swine flu?". Below the input field, a character count shows "3961 characters left" and a "Refine Question" button. The "About the event:" section provides a link to a BBC News article from April 2009. The "Related results from the QA repository" section features a link titled "Is there a connection between ozone levels and swine flu?" with a subtext "There may be a climate change link" and two buttons: "Accept answer and blog" and "Accept answer without blogging". The "Community Experts (excluding yourself)" section lists three experts: "Mr Michael Carlin", "GS-12 Ken c Nitz Capt.", and "Data Steward", with an "Ask Community" button. At the bottom, a note suggests forwarding the question to other peers.

The user may choose to accept one of the answers or refine the question and resubmit it to SKIPAL. Then the user must either refer the question to a community expert as identified by SKIPAL, or forward the question to another SKIPAL user (Figure 12).

Figure 12. Forward a question

UNCLASSIFIED

Maj Frank Gas's SKIPAL

Forward a Question to other SKIPAL Users

Forward the following question:

How long will it take to get through this traffic?

About the following event:

160600Z Apr 2009 (U) I-15 SOUTHBOUND: JAMFACTOR 0.8

To the following users:

- SKIPAL A.I.
- Joe Blow
- Michael Carlin
- Mr Skiweb Tester
- Edward Lai
- Mr Traffic Boy
- Data Steward

Select recipients:

Reset Submit

UNCLASSIFIED

Clicking on the **View Questions and Answers** button on the Enhanced Event page displays a new page (Figure 13) containing a list of all the questions that have been asked on that event. Clicking on a link under the Question column displays the Answer a Question page. Clicking on a link under the Action column displays the View Answers to a Question page (Figure 14) for that event. Clicking on the **Answer this question** link displays the Answer a Question page (Figure 18) for that event.


Figure 13. Questions page

UNCLASSIFIED

Maj Frank Gas's SKIPAL

Questions

About the following event

301202Z Apr 2009 (U) BBC News | News Front Page | World Edition : UK to see 'more' swine flu cases  [1 blog]

Date Asked	Question	Asked By	Action
301523Z Apr 2009	Is it time to start panicing???	Michael Carlin CIV SPAWARSYSCEN	1 answer Answer this question.

UNCLASSIFIED

Figure 14. View Answers to a Question page

UNCLASSIFIED

Maj Frank Gas's SKIPAL

View Answers to a Question

Joe Politics KTR zzzz

asked the following question:
How high will oil prices rise?

about the following event:
092021Z Mar 2009 (U) Yahoo! News: Science News : Iraq: Minister says OPEC aims to raise oil prices (AP)

[Answer this question.](#)

Answers ⁽¹⁾

2009-03-12 23:09:59.27 The price of gas will be \$4 per gallon by this summer.

Frank Gas
USSTRATCOM/J012

4.4 ALL ACTIVE EVENTS PAGE

The All Active Events page (Figure 15) lists all of the events that have not yet expired. It is nearly identical to the Recommended Events page, except for the Score column and the red **X** button used to filter events. It also has a PAL column, which will display an icon for those events that exceed the MEDL score threshold set in the user's profile page. The presence of the PAL icon means that this is a recommended event. By default, the All Active Events page sorts the events according to the event date, in reverse chronological order.

Figure 15. SKIPAL All Active Events page

UNCLASSIFIED

Maj Frank Gas's SKIPAL

[Update Profile](#)

Recommendations All Active Events Summary Q & A My Event Log

SKIPAL All Active Events

Results 1 - 30 of about 108 1 2 3 4 Next Table view: Normal Compact

Date	Title	Blog	Question	Categories	PAL
● 162207Z Mar 2009	(U) Predicted Conjunction Between BADR-B vs FENGYUN 1C DEB 16/2207Z Mar 09 (Miss Distance is Now 0.138 KM) Participants: GAPS SOCRATES			Space Ops	  

4.5 DAILY SUMMARY PAGE

The Daily Summary page (Figure 16) shows all of the SKIPAL events from the past 24 hours grouped by category. An event may appear in more than one category if the classifier assigns it to multiple categories. The Daily Summary page is a feature specifically requested by users. It allows them to rapidly scan the last 24 hours for events in certain categories. Clicking on the **triangle icon** to the left of each category name collapses/expands all of the events under that category. There are also **collapse all** and **expand all** buttons under the SKIPAL Daily Summary. The number in parentheses to the right of the category name shows the number of events that SKIPAL has identified for that category. If SKIPAL has not identified any events for that category, then that number will be zero and expanding the category will show NSTR (Nothing Significant to Report).

Figure 16. SKIPAL Daily Summary page

UNCLASSIFIED

Maj Frank Gas's SKIPAL [Update Profile](#)

Recommendations All Active Events Summary Q & A My Event Log

SKIPAL Daily Summary

[Collapse all](#) [Expand all](#)

▼ **Afghanistan Events (2)**

- 100746Z Mar 2009 (U) Pakistan tribe agrees to hand over Taliban
- 090730Z Mar 2009 (U) Pakistani Troops Kill 15 Militants in Northwest

▼ **Africa Events (6)**

- 100947Z Mar 2009 (U) UN tries to fill gaps left by Darfur aid expulsion
- 100944Z Mar 2009 (U) Sudan weighs options over quashing war crime warrant
- 100557Z Mar 2009 (U) Four peacekeepers wounded in Darfur attack
- 091723Z Mar 2009 (U) Zimbabwe Prime Minister: Car Crash Was Accident
- 091630Z Mar 2009 (U) Sudan Frees President's Top Opponent
- 091457Z Mar 2009 (U) Plane Crashes Into Uganda's Lake Victoria Killing 11

▼ **Asia Events (1)**

- 100936Z Mar 2009 (U) Sri Lanka bombing kills 10, wounds cabinet minister

▼ **Business Events (0)**

NSTR

▼ **China Events (7)**

- 101149Z Mar 2009 (U) U.S. declines to sell F-16 fighter jets to Taiwan: MP (Reuters)
- 101054Z Mar 2009 (U) Dalai Lama blasts 'brutal crackdown' in Tibet
- 101042Z Mar 2009 (U) Pro-Tibetan rallies mark anniversary of uprising
- 100908Z Mar 2009 (U) Dalai Lama demands Tibet autonomy, mourns past
- 100832Z Mar 2009 (U) China says U.S. naval ship broke the law
- 091840Z Mar 2009 (U) Dalai Lama to demand Tibet autonomy, mourn past

4.6 QUESTIONS AND ANSWERS PAGE

The SKIPAL Q&A page (Figure 17) has three sections: Questions for you to answer, Questions you asked, and Questions you've already answered. When a SKIPAL user asks a question about an event and forwards the question to another user, that question will appear under the recipient's Questions for you to answer section. The recipient of the question may choose to answer or ignore the question. Clicking on the **Answer this question** link brings up the page shown in Figure 18. Clicking on the **Ignore this question** link will remove the question from the list. The next section, Questions you asked, lists all of the questions that the user asked with links to the SKIWeb events. If any SKIPAL user has answered the question, then there will be a link showing the number of answers for that question. Clicking on that link will allow the user to view the answers on the View Answers to a Question page (Figure 19). If the question has not been answered, then the asker may retract the question by clicking on the **Delete this question** link. This will remove the question from the Questions you asked section as well as from the Questions for you to answer section of all referred SKIPAL users. The final section, Questions you've already answered, shows all of the questions that the user has answered. The user can use the **Answer this question** link to respond to the other users as new answers get posted.

Figure 17. SKIPAL Questions and Answers page

UNCLASSIFIED

Maj Frank Gas's SKIPAL [Update Profile](#)

Recommendations All Active Events Summary Q & A My Event Log

SKIPAL Questions and Answers

Questions for you to answer

Date Asked	Question	Action
181742Z Mar 2009	Will this affect the price of oil?	Answer this question Ignore this question

Questions you asked

181938Z Mar 2009	What does this mean for the price of gas?	1 answer Answer this question
092211Z Mar 2009	How come GM and Chrysler need bailouts but Ford does not?	Answer this question Delete this question

Questions you've already answered

182052Z Mar 2009	What does this mean for the price of gas?	1 answer Answer this question
122309Z Mar 2009	How high will oil prices rise?	1 answer Answer this question

UNCLASSIFIED

Figure 18. Answer a Question

UNCLASSIFIED

Maj Frank Gas's SKIPAL

Answer a Question

Answers (0)

Be the first to answer this question.

You are responding to the following question:
Will this affect the price of oil?

Asked by:
Edward Lai SPAWARSYSCEN

about the following event:
[Russia Calls to Rearm Military, Lashes Out at U.S.](#)

Enter your answer here and click the Submit button below.

4000 characters left

Figure 19. View Answers to a Question You Asked



As new answers appear, the question asker may share an answer with the SKIWeb community by allowing SKIPAL to blog it on SKIWeb or simply accept an answer without blogging. Only the user who asked the question will see the **Accept answer and blog** and **Accept answer without blogging** buttons. If the buttons are grayed out, then the asker has already accepted that answer. When the asker accepts an answer by clicking on either of the two buttons, SKIPAL adds the question and answer pair to its repository. SKIPAL will use this knowledge to suggest question and answer pairs to similar questions that are asked in the future.

4.7 RECATEGORIZE EVENT PAGE

From the Enhanced Event page, clicking on the **Recategorize Event** button brings up the Recategorize Event page (Figure 20). The user can choose up to three categories for each event.

Unlike recommendations, the categories model is shared. All users train the same classifier and see the same categories on the events.

Figure 20. Recategorize Event page

UNCLASSIFIED

Recategorize Event

The event you are recategorizing:

160702Z Mar 2009 (U) OPEC holds off from oil output cut (AFP)

Please select up to 3 event categories by holding down the SHIFT or CTRL key and then click Submit.

- Afghanistan
- Africa
- Asia
- Business
- China
- Congo
- Crime
- Disaster
- Economy
- Entertainment

Select categories:

Submit

History

171920Z Mar 2009 Frank Gas recategorized event from [Traffic] to [Business]

UNCLASSIFIED

4.8 USER PROFILE PAGE

The main SKIPAL page contains the five tabs. Clicking on the **Update Profile** link located in the upper-right corner of the main SKIPAL page displays the User Profile page (Figure 21). On the left side of the page, users can select from a list of predefined topics or create five additional topics. Users may enter more than one keyword (separated by a space) on the same line for each of the additional topics. In addition, users can specify interest in a geographic area by enabling a checkbox next to one or more regions in the list. SKIPAL will use the topics and geographic areas of interest to tailor each user's recommended events list.

Figure 21. SKIPAL User Profile page

UNCLASSIFIED

Maj Frank Gas's SKIPAL Profile

Details

Enter your command:

Enter your J-Code:
Please enter your whole code (e.g. J-323 vice just J-3).

Enter your job title:

Select your topics of interest:

- C4 / Net Ops
- Combating WMD
- Conjunction
- Cyber
- Deterrence Ops
- EXORD
- FRAGO
- Force Readiness
- Global ISR
- Global Strike Ops
- IMD Ops
- Information Ops
- Intelligence
- Missile Defense
- Network Warfare
- Nuclear
- OPORD
- OPSCAP
- PLANORD
- SC / Information Strategy
- Space Ops
- WARNORD

Enter up to 5 additional topics of interest:

Topic 1:

Topic 2:

Topic 3:

Topic 4:

Topic 5:

Select your geographic areas of interest:

- Afghanistan
- China
- HOA
- Iran
- Iraq
- Israel
- Lebanon
- North Korea
- Pakistan
- Philippines
- Russia
- Sudan
- Syria
- Turkey
- Venezuela

Settings

Event display mode: Exercise Real world Both

Recommendation threshold:
(between 0.0 and 1.0)

Number of similar events:
(no more than 25)

Number of suggested contributors:
(no more than 25)

Do you want to participate in the survey? Yes No

Do you want to reset your recommendation engine to its initial state? Yes No
This action cannot be undone once this form is submitted.

UNCLASSIFIED

On the right side of the page, users can select to see exercise events, real-world events, or both types of events in their recommended events list. The default value is **Both**. The recommendation threshold value is used with the values under the Score column on the recommended events tab to filter events. Only events with a score higher than the set threshold value will be recommended. All other events will be hidden. The default value is 0.45 but the recommendation threshold can be set to

any value between zero and 1.0. Each user should experiment with this setting to determine what value works best. The number of similar events and suggested contributors correspond to the size of those lists on the decorated event pages. The default value is 5 and cannot be set to a value greater than 25. “Do you want to participate in the survey?” is used to help new users train SKIPAL. By default, this value is set to **Yes** but the survey can be turned off at anytime by clicking on the **No** radio button. “Do you want to reset your recommendation engine to its initial state?” resets all of SKIPAL’s training for that user to a new user state and the process is irreversible. It is used when a user experiences unusual behavior in the recommended events list and SKIPAL is unable to demonstrate consistent learning from the user’s feedback. After making changes to the profile, the user clicks on the **Save changes and return to Recommendations** button at the bottom of the page.

5. RECOMMENDATION ENGINE EXPERIMENTS

SKIPAL employs several PAL technologies. However, we only conducted experiments and analysis with the recommendation engine to objectively measure PAL performance.

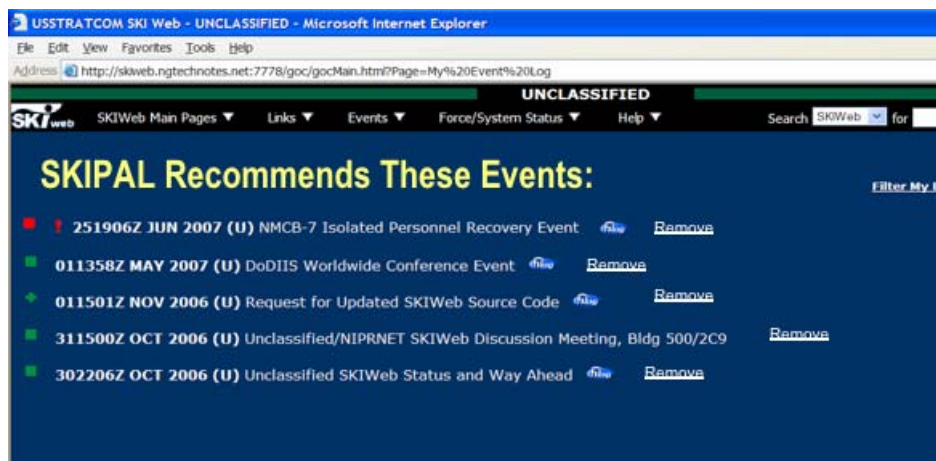
5.1 APRIL 2008 EXPERIMENTS (SPIRAL 2.1)

From 9 April through 2 May 2008, we conducted a set of experiments with 16 users and a modified version of SKIWeb running on a server in the USSTRATCOM EPL. This version of SKIWeb had a custom “SKIPAL Recommends” event list, which displayed events in a specified order for each user.

To set up the experiment, SKIPAL was preloaded with approximately 3 months worth of recent events, blogs, and read-by data. Thus, the recommendation engine had a head start on an interest model for most users. Some users were not active SKIWeb users prior to the experiment and had little or no history.

Every day during the experiment, a CD containing new data from the production SKIWeb was created and imported into the local instance of SKIWeb. SKIPAL queried the local SKIWeb for the last 24 hours of events (72 hours on Mondays to cover weekends). SKIPAL updated the recommendation engine with those events and then recommended a list of events (in order of decreasing relevance) for each user. Lastly, the list was displayed to the user on the SKIPAL Recommends These Events page (Figure 22).

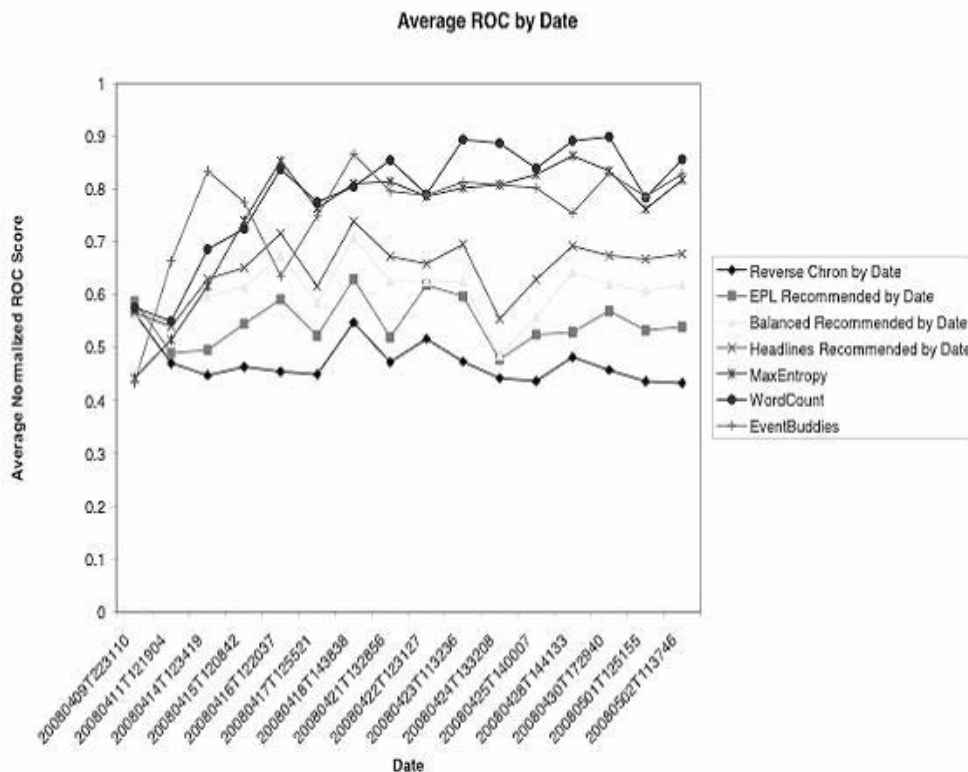
Figure 22. SKIPAL Recommends These Events page from April 2008 Experiment. Users indicated that an event was irrelevant to their interests by clicking the “Remove” link next to the event.



Each work day, our group of test users reviewed the list of events on the modified SKIWeb in the EPL. They were instructed to click the **Remove** button to remove any events irrelevant to them from the list.

Figure 23 shows the average Receiver Operator Characteristic (ROC) score compared to the reverse chronological order score from using iLink as the event recommendation engine for all of the users during the experiment. The performance of reverse chronological order hovers around 0.5, which is not better than a random recommendation engine.

Figure 23. Average ROC by date for the April 2008 Experiment. The “EPL Recommended by Date” reflects the performance of the iLink recommendation engine. The “Reverse Chron by Date” points were used as the control. They reflect the ROC score for events recommended in reverse chronological order, which is the default order of events in SKIWeb. A ROC score of 0.5 would be expected from a random recommendation engine.



Also shown are subsequent off-line tests we performed on the same data using candidate recommendation engines based on other PAL technologies.

As a recommendation engine, iLink did not perform as well as we had hoped. The averages in the chart show that iLink performed quite well with some users but poorly with others.

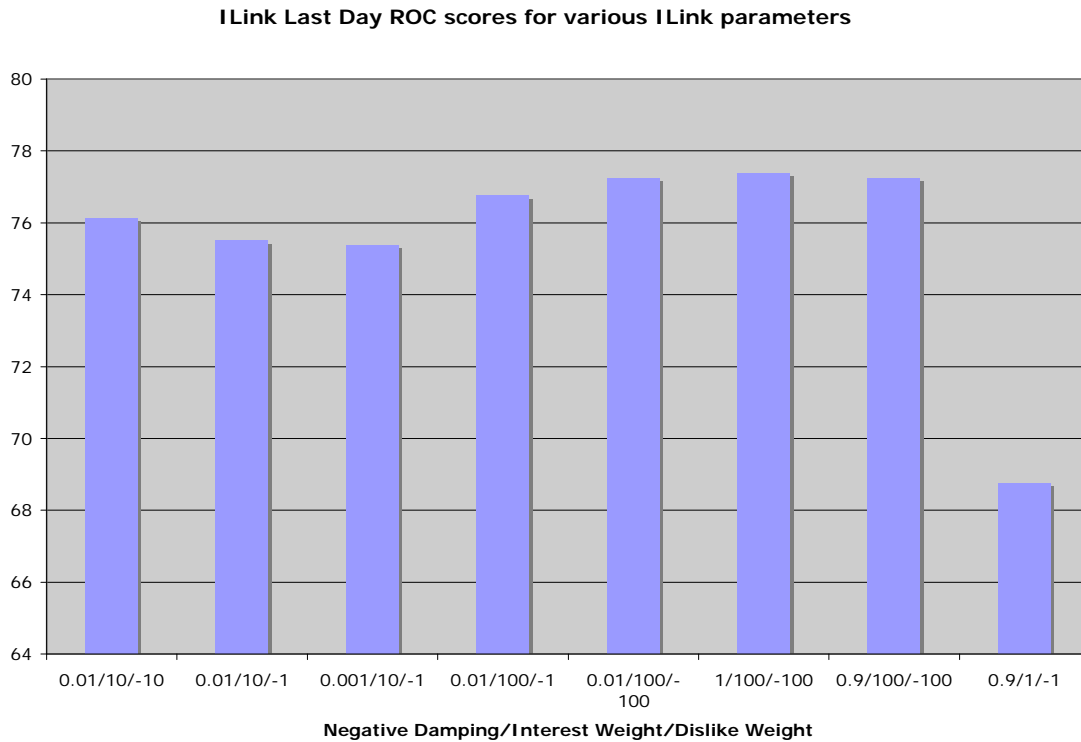
In general, recommendation engine architects face the challenge of integrating positive and negative feedback into a single model for interest. Historically, iLink has been used in an environment where positive signals tended to be much more important than negative signals. The developers theorized that iLink was essentially not giving much credence to the negative (user dislike) signals in SKIPAL. The negative damping factor and dislike expertise were identified as parameters that could be tuned. The value of the dislike parameter compared to the value of interest parameter (positive signal) could be interpreted as the learning rate. Therefore, increasing the magnitudes of these values would cause iLink to give more weight to these signals versus other signals in the model such as reading, authoring, or blogging.

The iLink team also theorized that iLink was getting lost in the noise of the event content. The team suggested that performance may improve if iLink considered only the headline of the event.

We performed offline experiments with these two variations and the results are shown as in Figure 23 as “Balanced” and “Headlines.” Both weighted positive and negative signals equally but only “Headlines” considered the title of the event alone. Performance improved in both cases so the developers’ theory about event headlines being more important was also brought to bear, but not to the point of surpassing the dedicated recommendation engines.

Figure 24. compares the performance of iLink with several parameter configurations. Note that the set of test users and the setup of the experiment itself were different. Therefore, the results in Figure 23 are not directly comparable to the results in Figure 24..

Figure 24. Effect of various parameter values on iLink’s performance as a recommendation engine for eight users on the last day of the April 2008 experiment. iLink was trained on data from the previous three months. The last column represents the approximate values used during the experiment.



Increasing the “dislike” value made a measurable difference. However, the best performance was obtained with a custom recommendation engine. Therefore, the development of MEDL began.

5.2 SPIRAL 2.1 DETAILED EVALUATION

5.2.1 Technical Approach

First, we analyzed the performance of the learning algorithm for each user. The survey results from Spiral 2.1 provided both position and relevance information for all of the events that were presented to each user. The relevance information is simply a yes/no response from each user as to the relevance of each event. We combined all of the survey results for each user and analyzed the

aggregate data. This was expected to have an averaging effect on the data, such that the performance of the learning algorithm would be less dependent on the specific events that were surveyed. Since we did not have knowledge of the true rankings of the events, we focused on the ability of the learning algorithm to assign relevant events to the top of the list and non-relevant events to the bottom of the list. Given the data we had to work with, we could only measure the learning algorithm’s ability to assign relative rankings as opposed to absolute rankings.

We needed a way to normalize the values for the positions across all of the days because the number of active events each user surveyed varied each day. For example, a user surveyed over 64 events on one day and over 179 events on another day. It would seem that the 53rd event in a list of 64 events would not have the same relevance as the 53rd event in a list of 179 events. Therefore, we normalized all of the positions to values between 1 and 100 according to the following formula:

$$\text{normalized position} = \left\lfloor \frac{\text{position}}{\text{total number of active events}} \times 100 \right\rfloor + 1.$$

Essentially, we used a process called “binning.” If the number of events was less than 100, then the data would be spread out across the bins. On the other hand, if the number of events was greater than 100, then there would be a grouping causing a loss of data because more than one position would be assigned to the same bin. One could use less than 100 bins but this would result in a greater loss of data.

After the positions for each surveyed event were normalized, we combined the data at each normalized position across the different survey dates for each user. The resulting data was then summarized by the two relative frequency distributions over the relevant and non-relevant events.

To facilitate the use of ROC analysis, we converted the rankings assigned to each event by the learning algorithm to binary classification labels. In other words, each normalized position was converted to a yes-or-no relevance rating, similar to the responses provided by the user in the survey results. By applying a threshold, all events with normalized positions equal to and above the threshold were classified as being *relevant*, while those with normalized positions below the threshold were classified as being *not relevant*. Such a threshold creates four possible outcomes: (1) if an event is relevant to the user and the learning algorithm classifies it as being relevant, then it is a *true positive (TP)*; (2) if an event is relevant to the user and the learning algorithm classifies it as being not relevant, then it is a *false negative (FN)*; (3) if the learning algorithm classifies an event that is not relevant to the user as being relevant, then it is a *false positive (FP)*; and finally, (4) if the learning algorithm classifies an event that is not relevant to the user as being not relevant, then it is a *true negative (TN)*. The decisions made by the classifier for a given threshold are summarized by the two-by-two confusion matrix in Figure 25. Note that different thresholds result in different confusion matrices (and, hence, different classifiers).

Figure 25. Confusion matrix for a given threshold

	Relevant to the user	Not relevant to the user
Classified as relevant	$a = \# \text{ of TPs}$	$b = \# \text{ of FPs}$
Classified as not relevant	$c = \# \text{ of FNs}$	$d = \# \text{ of TNs}$

Based on the entries in the confusion matrix, we could calculate several common metrics. The *true positive rate (TPR)* is equal to the probability that a relevant event is correctly classified as being relevant and is estimated as

$$TPR \approx \frac{a}{a + c},$$

where a and c are defined in Figure 25. Similarly, the *false positive rate (FPR)*, which is the probability that a non-relevant event is incorrectly classified as being relevant, is estimated as

$$FPR \approx \frac{b}{b + d}.$$

Two other important metrics include *recall*, which is equal to the true positive rate,

$$\text{Recall} = TPR,$$

and *precision*, which is estimated as

$$\text{Precision} \approx \frac{a}{a + b}.$$

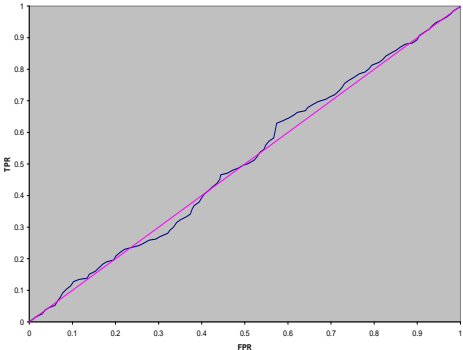
The ROC curve is given by the two-dimensional plot of the *TPR* versus the *FPR*. While a given threshold only corresponds to a single point in the ROC space, we can trace out the ROC curve by varying the threshold over all possible normalized position values and connecting the resulting points with straight lines. This is known as the *empirical or nonparametric* method for generating a ROC curve. The jagged appearance of the curve is due to the fact that the data is discrete instead of continuous. Note that the diagonal line connecting the two points (0, 0) and (1, 1) corresponds to a classifier that randomly guesses whether an event is relevant or not relevant. Thus, for those points in which the ROC curve falls below this diagonal line, the learning algorithm performs worse than random guessing.

The area under the ROC curve (AUC) is a scalar value that summarizes the expected performance of a classifier. Since we used the *empirical* method for generating the ROC curves, the area underneath the ROC curve can be decomposed into trapezoids and is easily calculated using the trapezoidal rule.

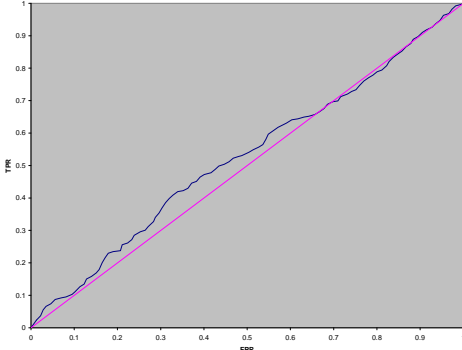
5.2.2 Spiral 2.1 Results

The ROC curves for all of the users are shown in Figure 26 Plotted along with the ROC curves is the diagonal curve, which corresponds to a random classifier.

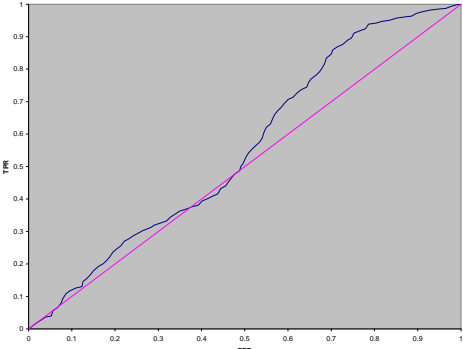
Figure 26. ROC curves for individual users in the Spiral 2.1 experiments. (Figure continued on following pages.)



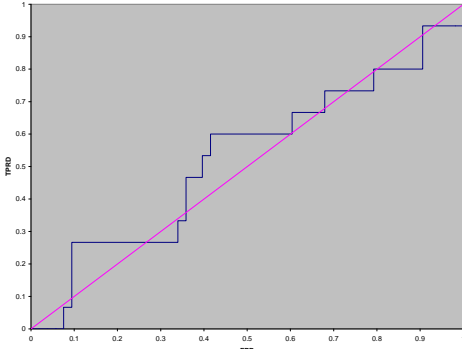
1276



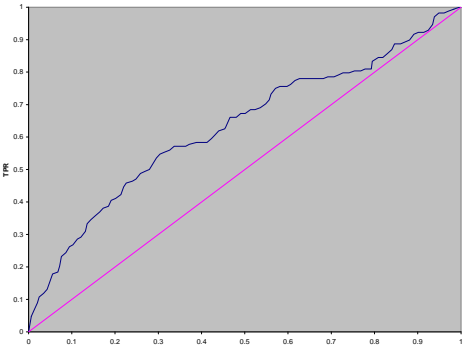
1691



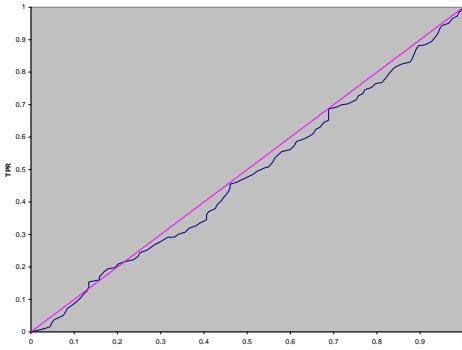
2250



2347

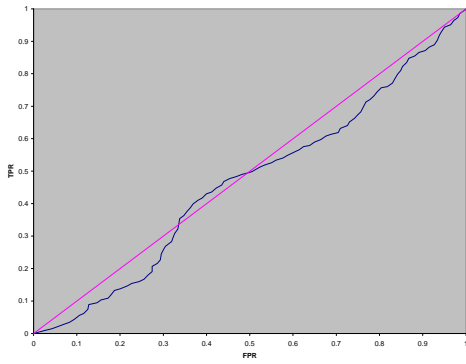


2372

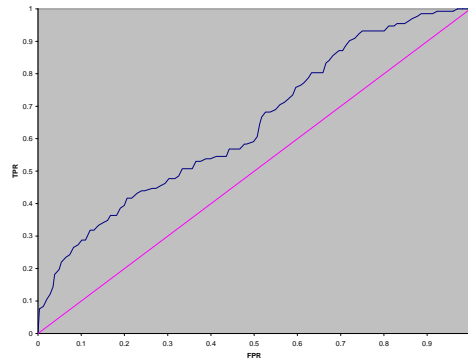


6070

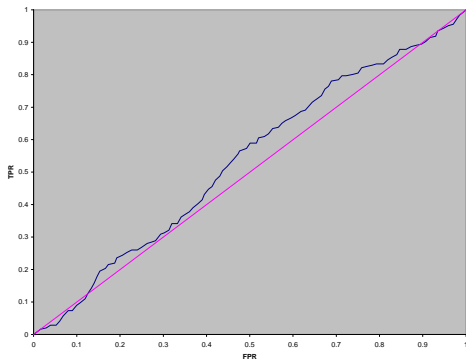
Figure 26. ROC curves for individual users in the Spiral 2.1 experiments. (Continued)



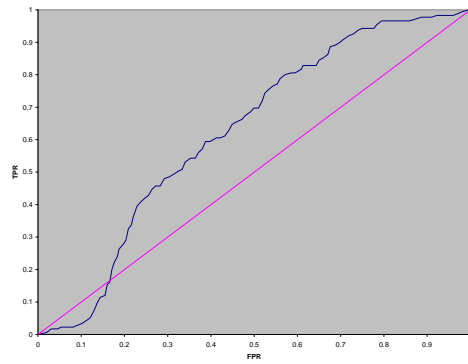
6389



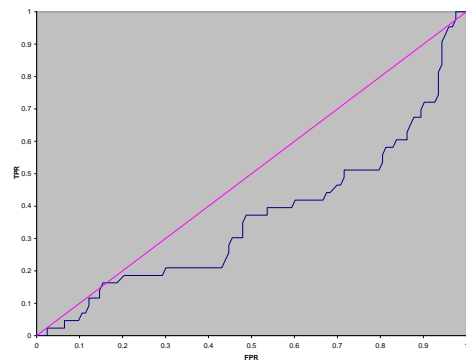
6983



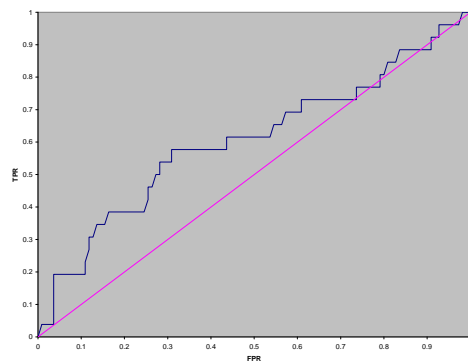
8615



10694

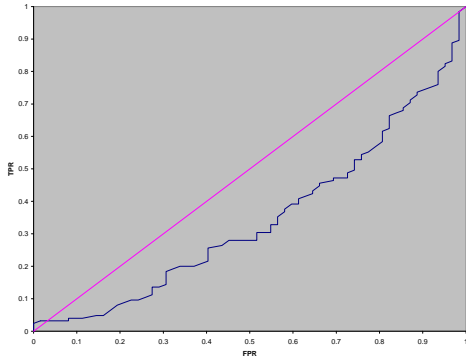


12353

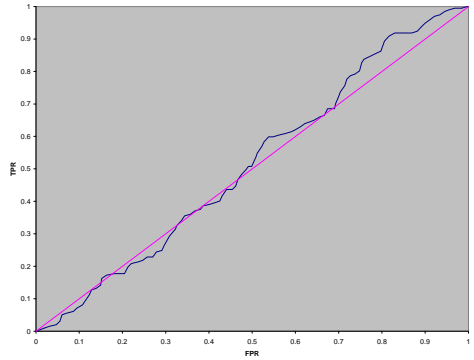


22974

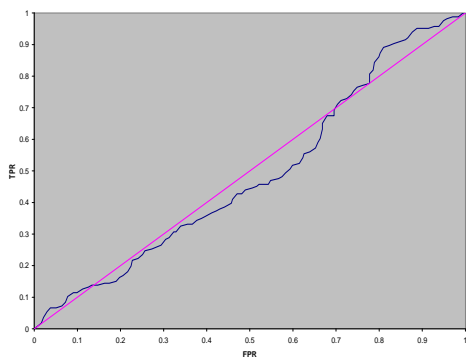
Figure 26. ROC curves for individual users in the Spiral 2.1 experiments. (Continued)



26287



39389



46766

The AUC values and the standard errors for all of the users are shown in Table 1.

Table 1. AUC values for Spiral 2.1.

User ID	AUC	Standard Error of AUC	# of Survey Responses
1276	0.506	0.018	995
1691	0.526	0.019	1095
2250	0.554	0.017	1137
2347	0.526	0.089	68
2372	0.631	0.026	655
6070	0.476	0.026	521
6389	0.466	0.019	974
6983	0.641	0.027	704
8615	0.534	0.022	859
10694	0.625	0.022	730
12353	0.364	0.052	166
22974	0.605	0.068	136
26287	0.346	0.042	187
39389	0.515	0.025	578
46766	0.488	0.027	540

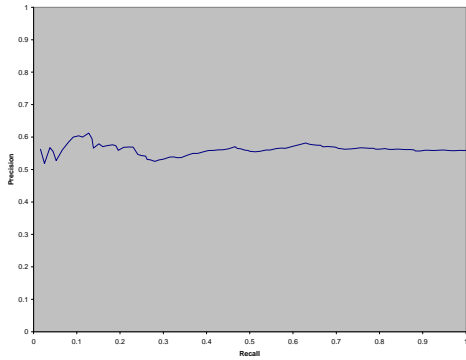
The standard errors are calculated using the formula given in reference [2]. More detailed steps of the use of this formula are given in reference [3]. Note that these standard errors tend to be lower for users with a higher number of survey responses, which makes sense. We can state with a high degree of confidence that the learning algorithm performed much better for users 2372 and 6983 than for any of the other users. However, the same cannot be said for user 22974. Even though the AUC value was also greater than 0.6, user 22974 has a much higher standard error than the two users previously mentioned. The ROC curves and AUC values also seem to indicate that the learning algorithm performed relatively poorly for users 12353 and 26287. As for the remaining users, their ROC curves straddled the diagonal line between (0, 0) and (1, 1), such that their AUC values were around 0.5. This means that the learning algorithm performed about the same as a random binary classifier.

A straightforward calculation of the average of the AUC values in Table 1 yields 0.520.

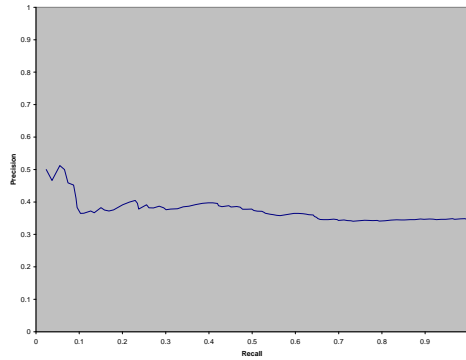
Note that such a calculation treats each AUC value equally and does not take into consideration their standard errors.

The precision-recall curves for all of the users are shown in Figure 27. We used linear interpolation to connect the points along the curve. While it has been argued that linear interpolation leads to overly optimistic results (see reference [4]), we claim there are enough points along the curve for most of the users to give accurate results.

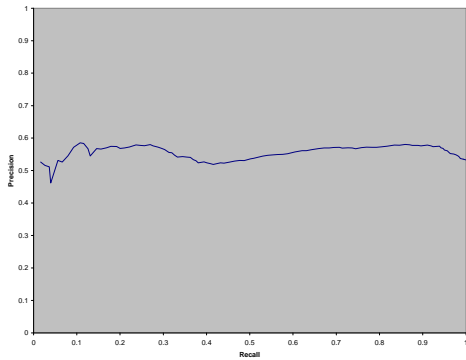
Figure 27. Precision and recall curves for users in the Spiral 2.1 experiment. (Figure continued on following pages.)



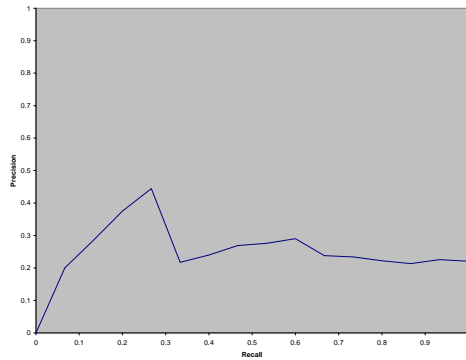
1276



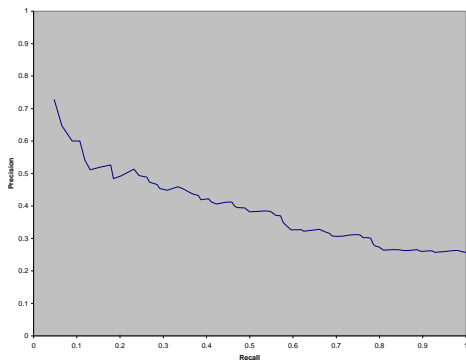
1691



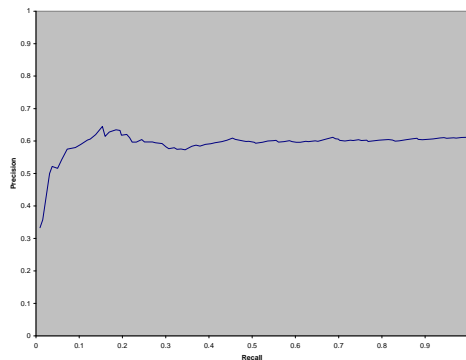
2250



2347

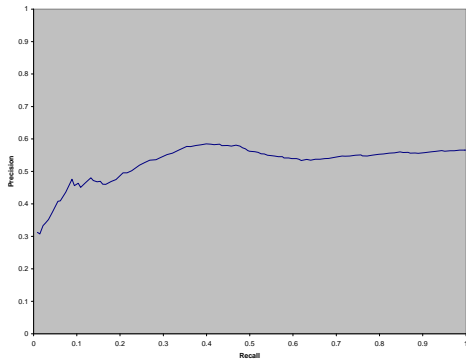


2372

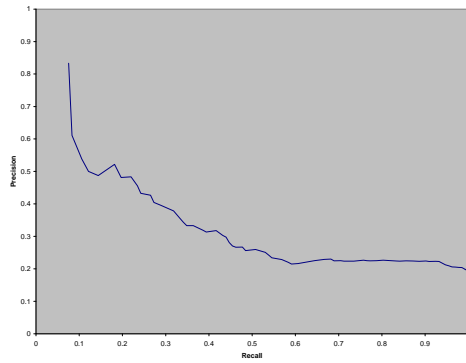


6070

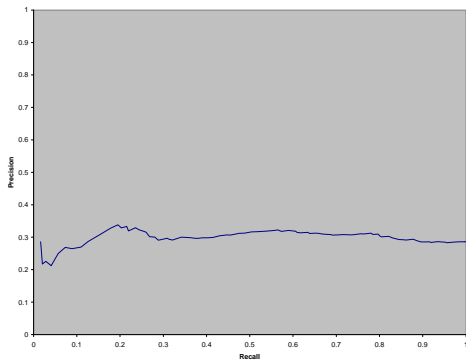
Figure 27. Precision and recall curves for users in the Spiral 2.1 experiment. (Continued)



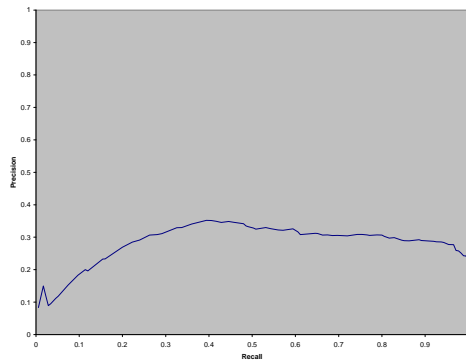
6389



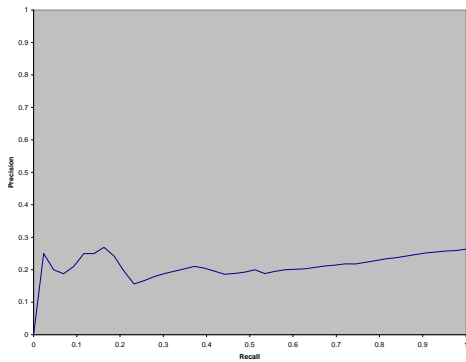
6983



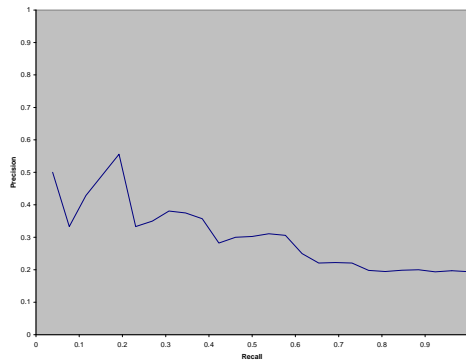
8615



10694

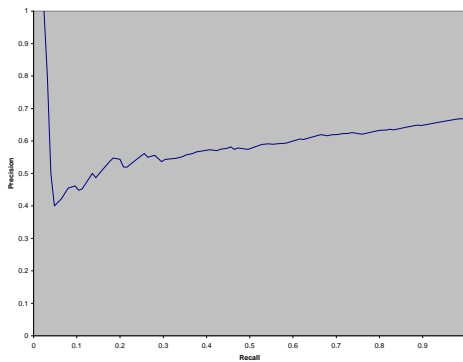


12353

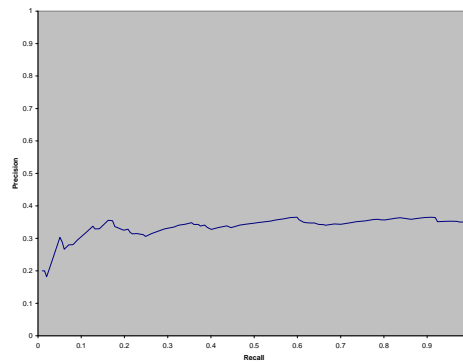


22974

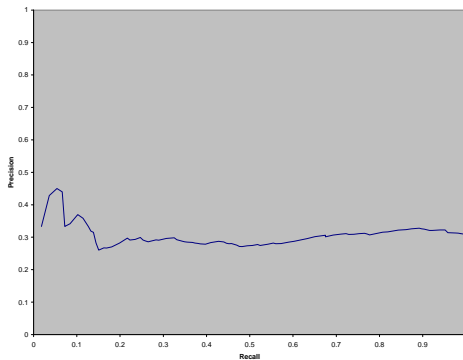
Figure 27. Precision and recall curves for users in the Spiral 2.1 experiment. (Continued)



26287



39389



46766

5.3 AUGUST AND NOVEMBER 2008 EXPERIMENTS (SPIRALS 2.3 AND 2.4)

These experiments were conducted at USSTRATCOM. The SKIPAL servers were located in the EPL. Earlier experiments required that the users go to the EPL to use SKIPAL. However, this time we obtained permission for users to connect to the server from their desktops. This made it easier for users to participate.

Users were instructed to use SKIPAL as they would normally use SKIWeb on a daily basis. We also educated them about the “thumbs-up” and “thumbs-down” functionality and described the other features of SKIPAL.

5.3.1 Survey Methodology

As described in the architecture section, the Survey page is displayed when the user first logs into SKIPAL and subsequently when the user visits the Recommendations page or the All Active Events page. Users can opt out of the survey by checking a box on their profile page. But since we explained the purpose of the experiment in advance, nearly everyone in the user test group participated in the surveys.

We frequently solicited feedback on the users’ experience during both experiments and received both verbal and written comments.

5.3.2 Analyses

5.3.2.1 Sampled Precision and Recall Analysis

Our survey algorithm presented the user with an event that would not normally be recommended 2/3 of the time. For example, given a total of 100 active events, the recommendation engine might recommend 10 of those events because the user limited the display to 10 items. Or, in the most recent instance of SKIPAL, the user set the score threshold such that only 10 events exceeded the recommendation threshold. As a result, the 10 recommended events have a 1/3 probability and the remaining 90 events have a 2/3 probability of being shown to the user. We call this “the 1/3-2/3 rule.”

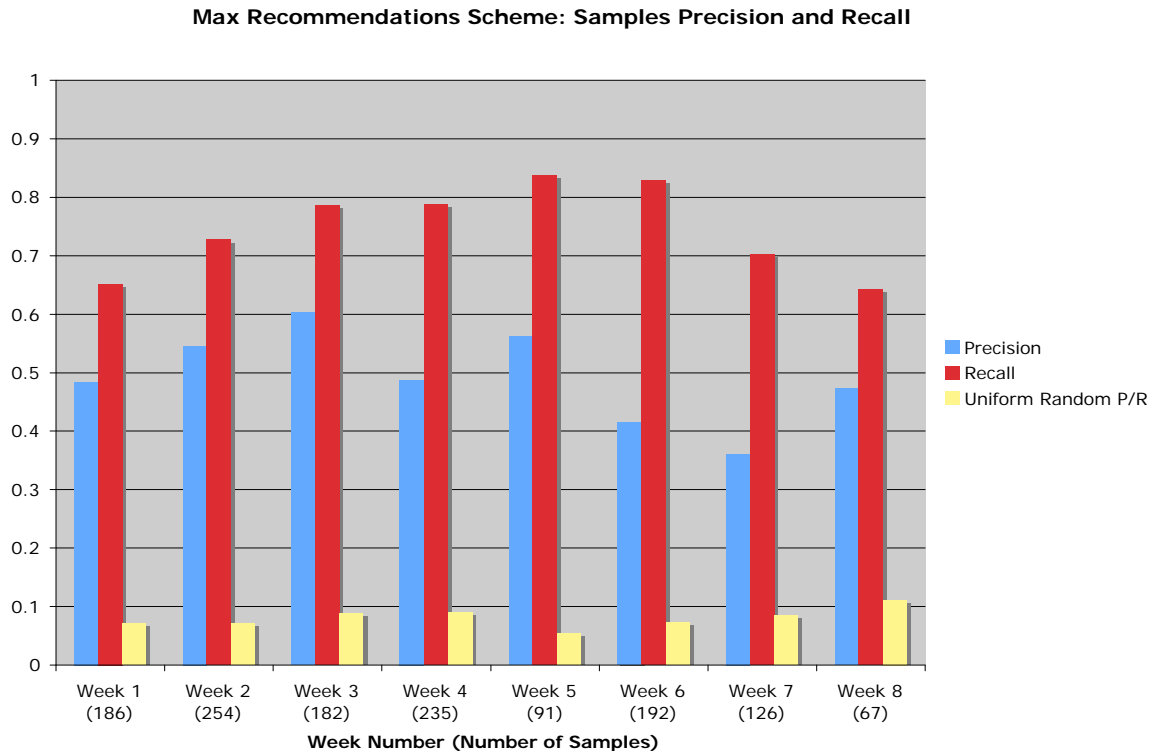
We recover an estimate of precision and recall from the data as follows. Precision is defined as the fraction of recommended items that were relevant to the user. Intuitively, it is a measure of the quality of the recommendations. Therefore, we only consider samples taken from the recommended window. The fraction of these samples that a user said was relevant represents the estimated precision.

Recall represents the fraction of all relevant items that were recommended. Recovering the recall is not as simple as counting the number of relevant events (exceeds the recommendation threshold) inside the recommendation window and dividing by the total number of relevant items because of the 1/3-2/3 rule. In addition, our sampling was biased to select events outside the window twice as often as those inside. Therefore, we removed the bias by adjusting the weight of a relevant event outside the window as 0.5 instead of 1.0 when counting.

It is easy to see that this approach works on a toy problem. Suppose in the preceding example that there is a total of 20 relevant events, 10 of which appear inside the recommendation window of size 10. Clearly, the recall is 0.5 (10 of the 20 relevant events were recommended). However, our sampling scheme will return events outside of that window twice as often. If we simply divided the number of relevant samples inside the window by the total number of relevant samples, we would estimate a recall of 1/3. Halving the count of relevant samples outside of the window properly compensates for our sampling bias.

Using this scheme, we can estimate the precision and recall achieved during the August–October 2008 (Spiral 2.3) and November–February (Spiral 2.4) experiments. During the August–October experiments, the user was able to set a maximum number of recommendations to display. This number sets the window size for a given survey sample. Figure 28 shows the estimated precision and recall for survey samples accumulated over all test subjects for each week, beginning on 27 August. In the course of interpreting these results, note that precision and recall metrics are tightly coupled. A large recommendation window is more likely to result in a high recall. Even a random recommendation engine will have a recall of 1.0 if the number of events is slightly less than or equal to the window size but the precision will suffer. On the other hand, if the recommendation engine is good at modeling one particular aspect of the user’s interests and enough events of that type are present, then having a small window will produce a high precision but recall will be poor if the number of relevant events is smaller than the recommendation window. Ideally, knowing the number of relevant items in advance would easily determine the recommendation window. But in the real world, different users have different interpretations of what is relevant to them and, therefore, it is difficult to identify a particular window size that works for everyone.

Figure 28. Sampled precision and recall for the “Max Recommendations” survey data set, representing dates from 27 August 2008 through 21 October 2008. The Expected P/R values represent the expected precision and recall assuming the recommended items were randomly ordered (i.e., the relevant items are uniformly distributed throughout the list).



Whether precision or recall is more important is a subjective discussion. Some users prefer to see all of the events and sift through the data to find the relevant ones. Those users might prefer a higher recall to precision. Other users might not be as tolerant to “noise” in their recommended events and will value precision more highly.

Here is an interesting value for comparison, which we call the “Uniform Random P/R.” This is the expected precision and recall value for a uniform random recommendation engine. For such an engine, we expect the relevant events to be distributed uniformly throughout the list. Therefore, the expected precision and recall will be equal to the window size divided by the total number of active events at the time of the survey. Comparing the precision and recall of the MEDL recommendation engine to this value provides a valuable indication of how well MEDL does at putting relevant items into the recommendation window. For example, given the precision for week 5 in Figure 28 we can say that the density of relevant events in the recommendation window was approximately 10 times higher for MEDL and recall was about 16 times higher than for a random recommendation engine.

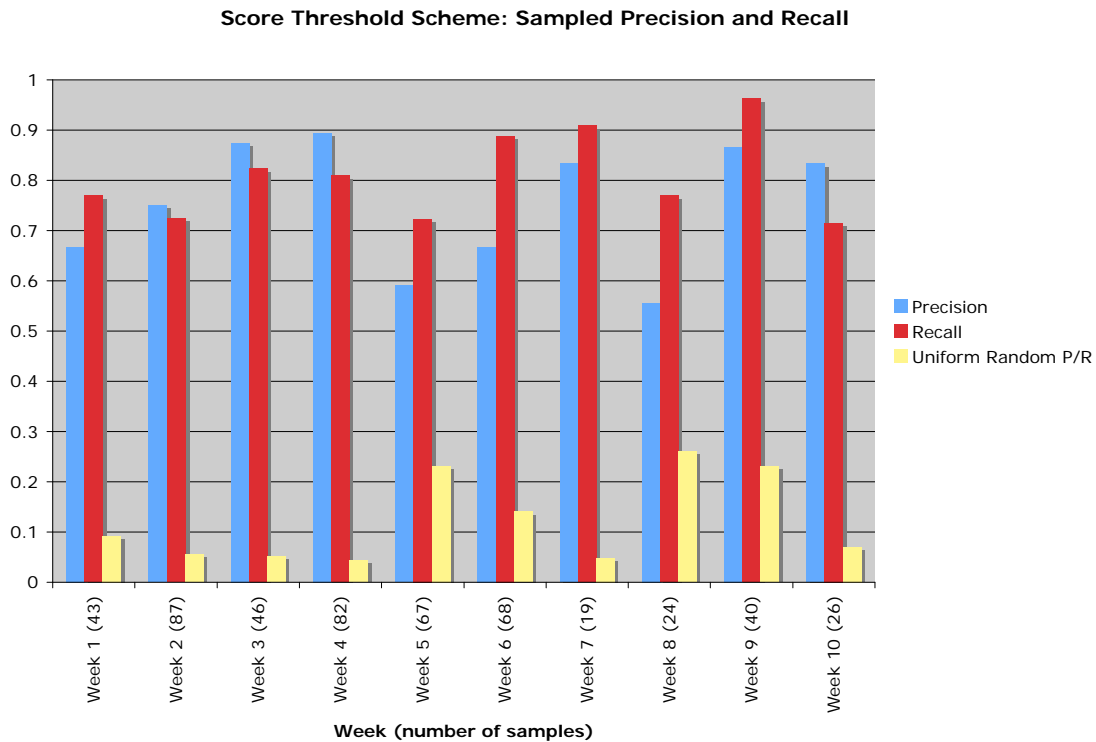
However, the roughly flat trend in the experiments does not lead us to conclude much about MEDL’s learning rate. MEDL starts with no idea of the user’s interest until the user trains it. MEDL takes into account information about each user’s actions such as events created, blogged, bookmarked, or read. In some cases, a single user only generated a few survey data points in a single day or even a week. Thus, it was difficult to rely on short-term performance estimates. But after we

aggregated all of the users for each week, there were enough data points to draw meaningful conclusions.

The next experiment began on 7 November 2008 and ran for approximately 10 weeks. Spiral 2.4 of SKIPAL was deployed with the MEDL scores exposed and the number of events displayed on the Recommendations page was driven by a user-specified score threshold. Users could lower the threshold to allow more events to be displayed, or raise the threshold to filter out more events. The results of this experiment, in terms of sampled precision and recall, are shown in Figure 29. The MEDL user data were *not* reset at the beginning of the experiment.

These results reflect a higher precision and recall than the previous experiment but the number of samples is smaller because we had less user participation. Nevertheless, these results are promising and suggest that MEDL is learning to do a good job at determining what is valuable to SKIWeb users.

Figure 29. Sampled precision and recall for the “Score Threshold” survey data set, representing dates from 7 November 2008 through 15 January 2009. The Expected P/R values represent the expected precision and recall assuming the recommended items were randomly ordered (i.e., the relevant items are uniformly distributed throughout the list).



5.3.2.2 Subjective Feedback

Feedback from users on the MEDL recommendation engine was mostly positive. With only a few exceptions, the recommendation engine quickly converged onto a set of relevant events for the user. The exceptions usually indicated a bug in SKIPAL or unbalanced feedback from the user, such as using the “thumbs-down” button much more often than “thumbs up,” or using “thumbs down” instead of the “X” to remove irrelevant events.

With the most recent version of SKIPAL, users expect good intuitive performance from the recommendation engine, and have shifted their criticisms to other areas of the software application (since addressed, or noted elsewhere in this report).

5.4 SPIRAL 2.3 EVALUATION

The most notable difference between the Spiral 2.1 and the Spiral 2.3 survey results was that in Spiral 2.3, we obtained relevance knowledge for only one of the events in the list of active events instead of all of the events. Given that we had less feedback data to work with, we had to combine the data from all of the surveys for each user. Thus, we used the same approach as in Spiral 2.1. That is, we normalized all of the position values to the range of integers from 1 to 100 according to the total number of active events at the time the user was surveyed. We employed the empirical method for generating the ROC curves and calculating the AUC. We expected our estimates of the true ROC curves and AUC values to be less accurate than in Spiral 2.1.

5.4.1 Spiral 2.3 Results

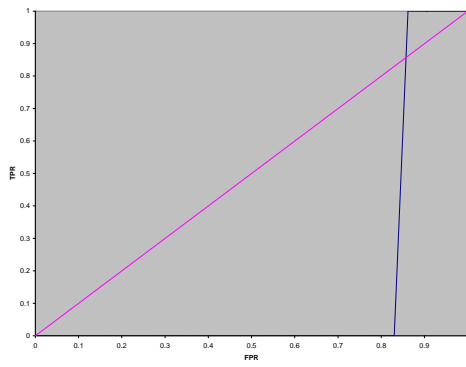
The ROC curves for all of the Spiral 2.3 users are given in Figure 30.

The AUC values, their standard errors, and the number of survey responses for each user are given in Table 2.

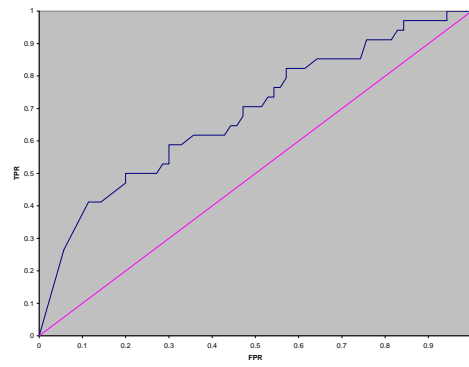
As expected, the ROC curves in Spiral 2.3 are more jagged. In general, the standard errors for the AUC values are much higher than in Spiral 2.1. Nevertheless, we can say with a high degree of confidence that the learning algorithm performed very well for users 1691, 1964, 6110, and 1049457, and very poorly for user 1080. Note that user 1080 was a test user and was not officially part of the experiment. Note that the results for users 5485, 7569, 8615, 36003, and 1053616 should be ignored due to lack of data. User 5485 had a single “yes” and “no” response, while user 1053616 only had “no” responses. Compared to the values in Table 1, the AUC values in Spiral 2.3 are distributed over a much larger range.

Ignoring the results for the users with too little data, we calculated an average AUC value of 0.669. Again, this value does not take into account the standard errors of the AUC values.

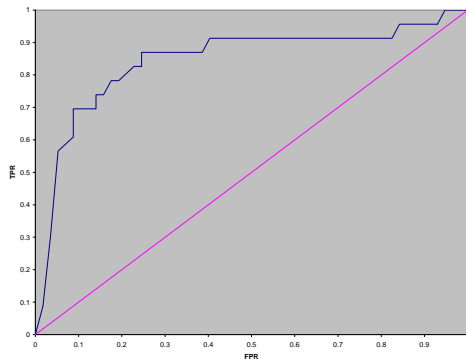
Figure 30. ROC curves for Spiral 2.3 users. (Figure continued on following pages.)



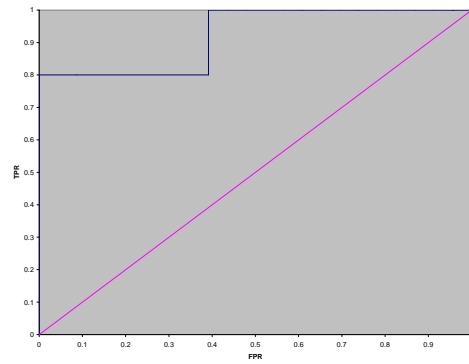
1080



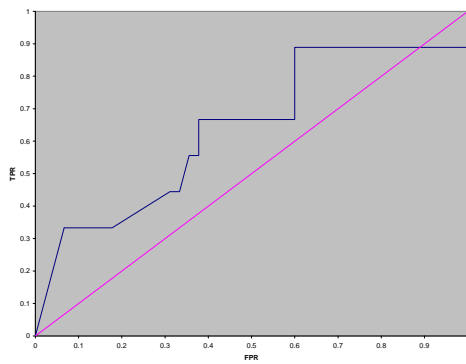
1460



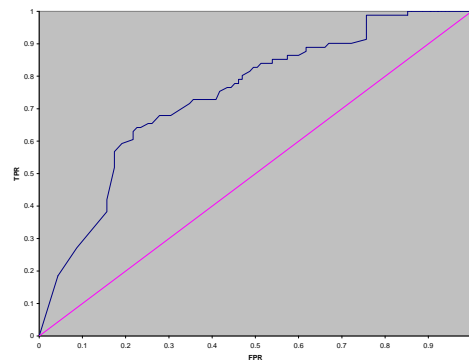
1691



1964

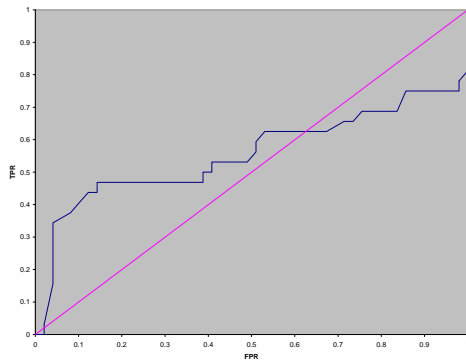


2004

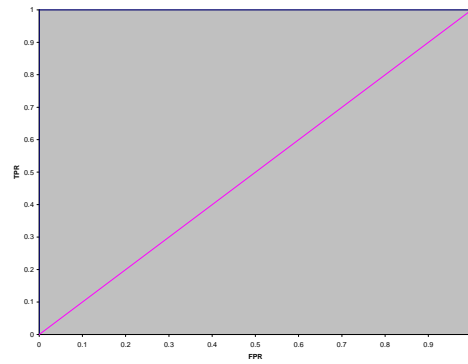


2250

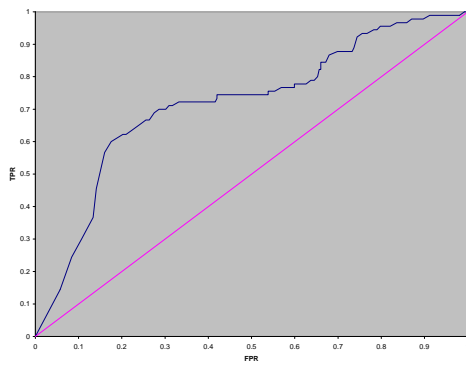
Figure 30. ROC curves for Spiral 2.3 users. (Continued)



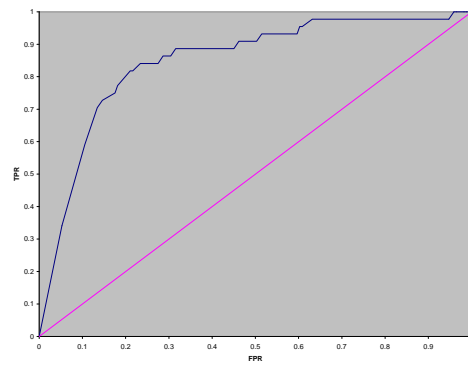
2372



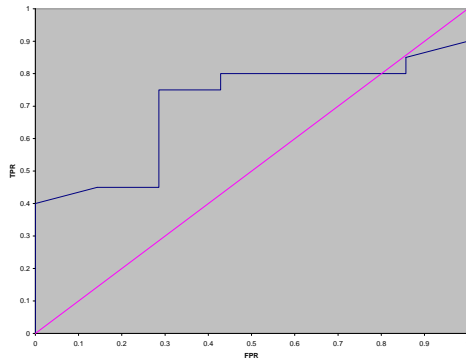
5485



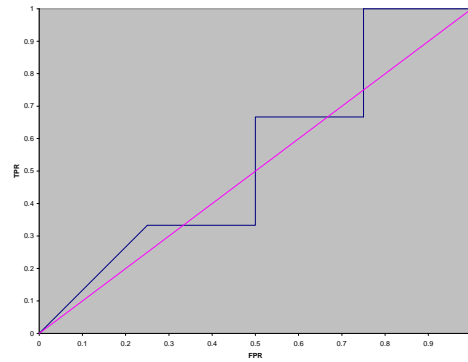
6070



6110

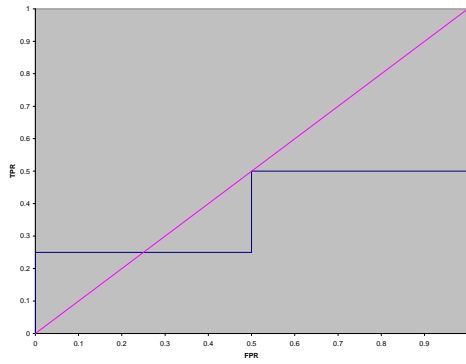


6973

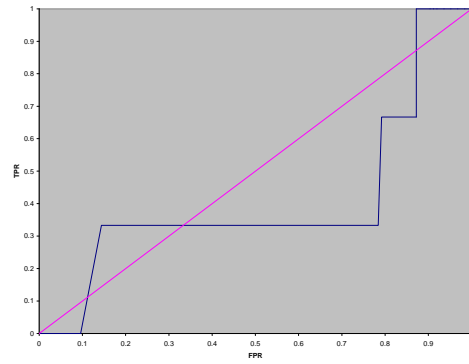


7569

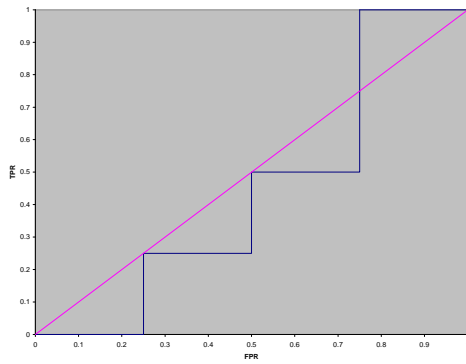
Figure 30. ROC curves for Spiral 2.3 users. (Continued)



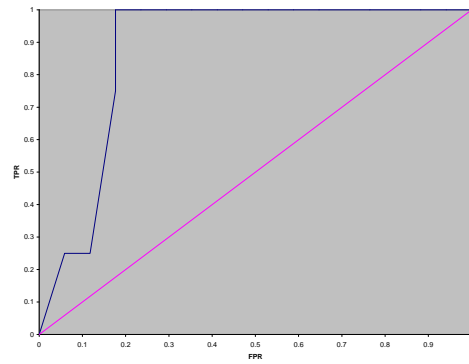
8615



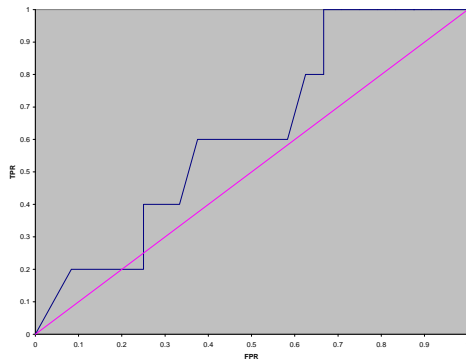
24384



36003



1049457



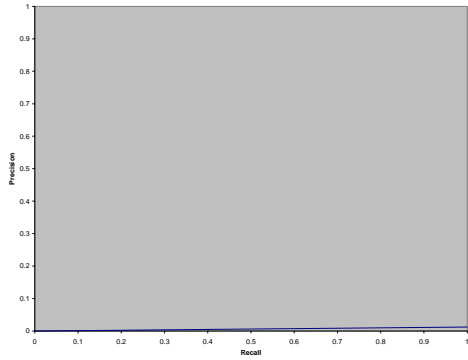
1053565

Table 2. AUC values for Spiral 2.3.

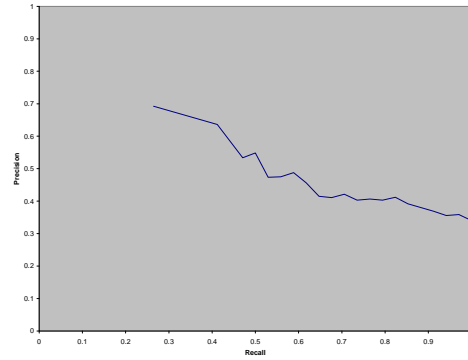
User ID	AUC	Standard Error of AUC	# of Survey Responses
1080	0.155	0.038	95
1460	0.684	0.057	104
1691	0.844	0.056	80
1964	0.922	0.075	28
2004	0.637	0.110	54
2250	0.741	0.036	196
2372	0.555	0.075	81
5485	1	0	2
6070	0.718	0.033	352
6110	0.848	0.033	215
6973	0.7	0.104	27
7569	0.542	0.235	7
8615	0.375	0.238	6
24384	0.407	0.196	128
36003	0.438	0.222	8
1049457	0.875	0.077	21
1053565	0.617	0.125	29
1053616	-	-	7

The precision-recall curves for all of the users are given in Figure 31.

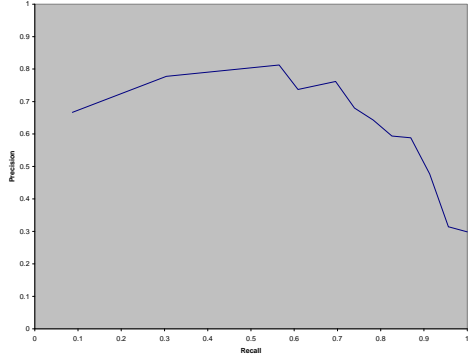
Figure 31. Precision-recall curves for the Spiral 2.3 (August 2008) experiments. (Figure continued on following pages.)



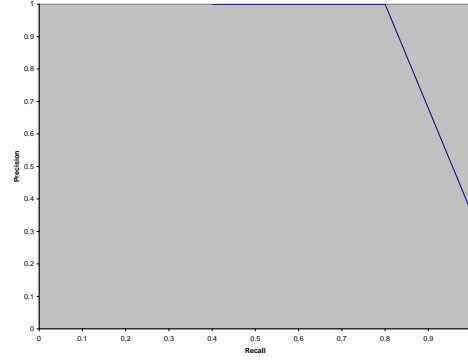
1080



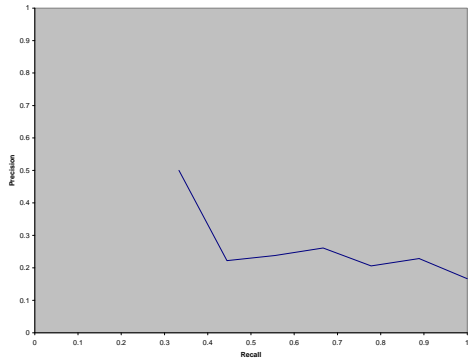
1460



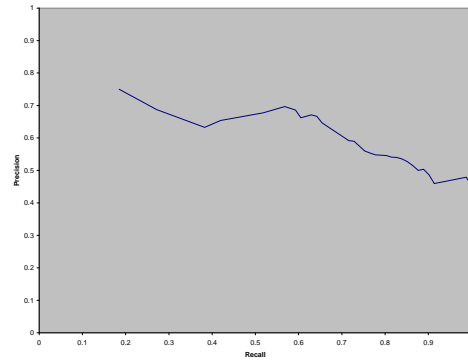
1691



1964

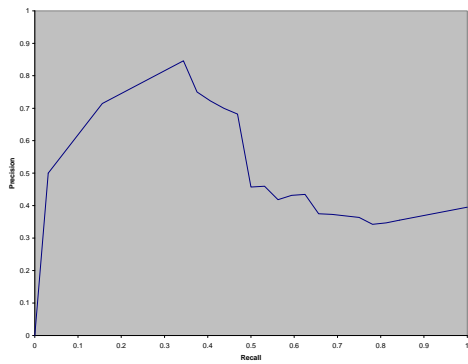


2004

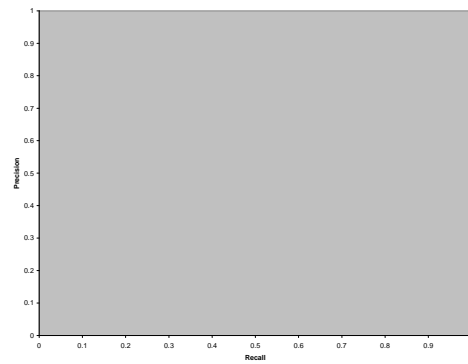


2250

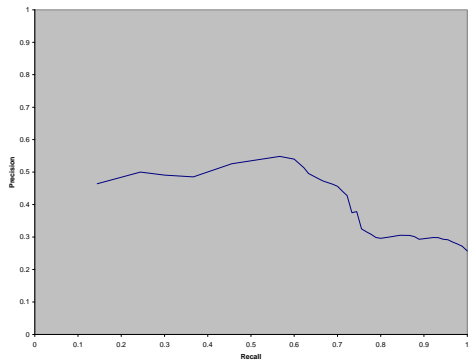
Figure 31. Precision-recall curves for the Spiral 2.3 (August 2008) experiments. (Continued)



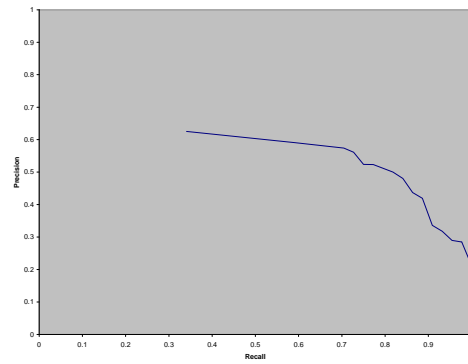
2372



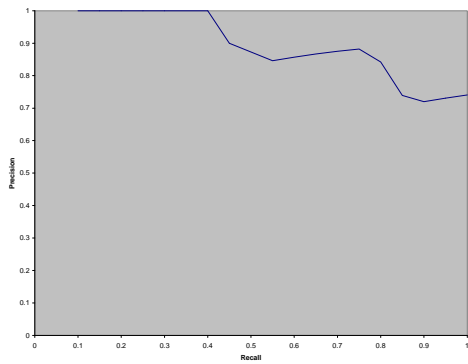
5485



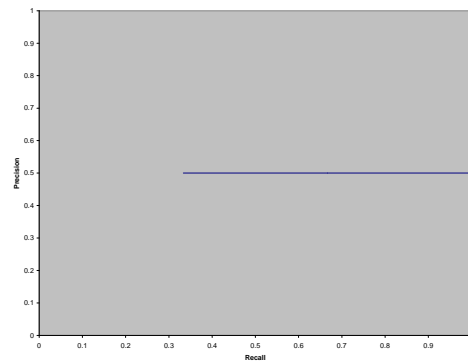
6070



6110

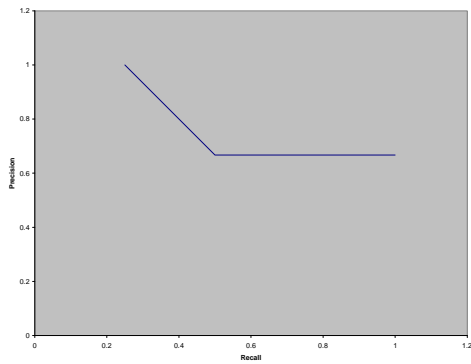


6973

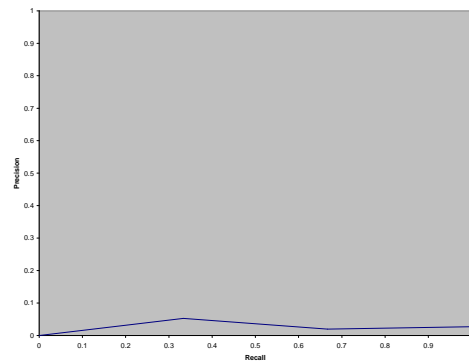


7569

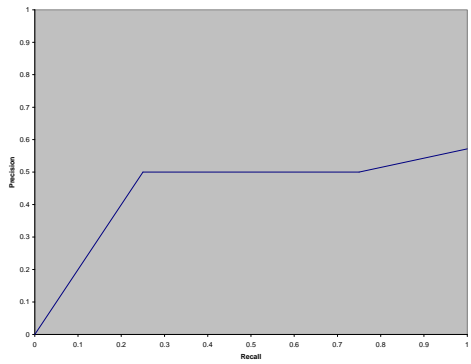
Figure 31. Precision-recall curves for the Spiral 2.3 (August 2008) experiments. (Continued)



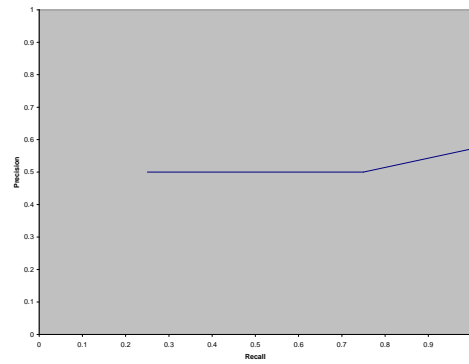
8615



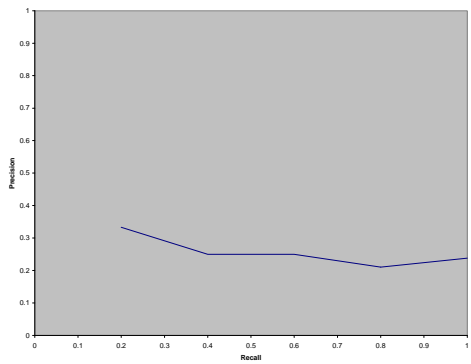
24384



36003



1049457



1053565

6. SUMMARY

6.1 CONCLUSIONS

When adding PAL technology to any system, it is easy to place too much faith in objective metrics, anecdotal praise, or criticism. We learned that if too much emphasis is placed on measuring performance, then the user experience can be a long and difficult process. On the other hand, if we focus on the user experience, we risk compromising our ability to accurately measure performance. We addressed this dilemma in SKIPAL Phase 1.

SKIPAL Phase 2 used the lessons learned from Phase 1 to avoid the same pitfalls. Significant new features and technologies such as a new recommendation engine, categorization engine, and Q&A interface were introduced and created in the context of a system that could be used by end-users in the course of their daily work.

In conclusion, based on these experiments, PAL provides a useful means of navigating the growing flood of SKIWeb information.

7. REFERENCES

1. Davitz, J., Yu, J., Basu, S., Gutelius, D., and Harris, A. 2007. "iLink: search and routing in social networks." In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Jose, California, 12–15 August 2007). KDD '07. ACM, New York, NY, 931-940. DOI=<http://doi.acm.org/10.1145/1281192.1281292>
2. Hanley, J. A. and McNeil, B. J. "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, pp. 29-36, Apr. 1982.
3. Wu, J. C. and Wilson, C. L. "Nonparametric Analysis of Fingerprint Data on Large Data Sets," *Pattern Recognition*, vol. 40, no. 9, pp. 2574-2584, Sep. 2007.
4. Davis, J. and Goadrich, M. "The Relationship Between Precision-Recall and ROC Curves," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 25–29 Jun 2006, pp. 233-240.

Approved for public release; distribution is unlimited.



SSC Pacific
San Diego, CA 921542-5001