15th ICCRTS
"The Evolution of C2"

# SKIPAL: The Incorporation of Machine Learning Technology into the Strategic Knowledge Integration Web

Topic(s)
Information Sharing and Collaboration Processes and Behaviors
Networks and Networking
C2 Architectures and Technologies

Douglas S. Lange, SPAWARSYSCEN Pacific
Edward C. Lai, SPAWARSYSCEN Pacific
Michael Carlin, SPAWARSYSCEN Pacific
Andrew S. Ling, SPAWARSYSCEN Pacific
Kellie Keifer, SRI International
Bill Deans, SRI International
Ken Nitz, SRI International
Laura Tam, SRI International
John Bolton, Northrop Grumman Corporation
Bill Graves, Northrop Grumman Corporation
Bill Reestman, Northrop Grumman Corporation

Point of Contact
Douglas S. Lange
Space and Naval Warfare Systems Center, Pacific
53570 Silvergate Ave Room 1432
San Diego, CA  92152-5147

(619)553-6534
doug.lange@navy.mil

# Abstract

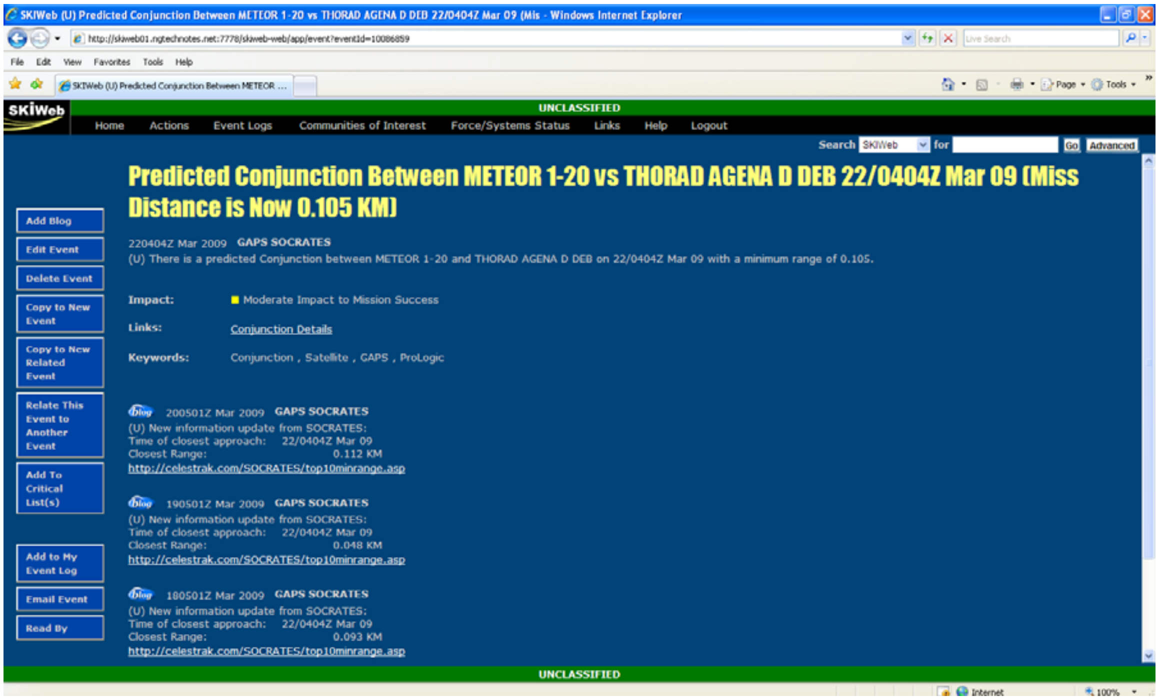
The Personalized Assistant that Learns (PAL) program was a DARPA research program with the primary goal of creating an integrated system that can adapt to changes in its environment and the users' goals and tasks without programming assistance or technical intervention. SKIWeb is an information aggregation system based at USSTRATCOM and is available to anyone on SIPRNET, a secure internet for the US Department of Defense. With a user base over twenty five thousand, and a constantly growing number of human and automated contributors, SKIWeb content threatens to overwhelm users, causing them to miss critical information amid a deluge of information. We hypothesized that PAL technology could be used to learn the information the user requires by observing the implicit and explicit signals in their interaction with SKIWeb; and that further, PAL technology could help with event identification and to expose the relationships between events and SKIWeb users, both of which could be leveraged to improve efficiency and quality in USSTRATCOM operations. Based on experimental results, USSTRATCOM has made the required budget requests to transition SKIPAL to a program of record. This paper describes the technologies incorporated into SKIPAL, results of the experimentation, and methods that have led to both a technical and programmatic transition success.

# Introduction

## SKIWeb

The Strategic Knowledge Integration Web (SKIWeb) is a web-based application for integrating and disseminating information on SIPRNET and maintaining situational awareness, based at the United States Strategic Command (USSTRATCOM). 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*. A portion of a rendered SKIWeb event is shown in **Error! Reference source not found.**.

Users and automated systems can annotate an event with a *blog*. Slightly different than the usual definition of a *weblog*, a SKIWeb blog is a short text description that is appended to an event. Blogs could include corrections, elaborations, questions and answers, or acknowledgements. The event in **Error! Reference source not found.** shows three blogs.


SKIWeb (U) Predicted Conjunction Between METEOR 1-20 vs THORAD AGENA D DEB 22/0404Z Mar 09 (Mis - Windows Internet Explorer
http://skiweb01.ngtechnotes.net:7778/skiweb-web/app/event?eventId=10086859
Live Search
File
Edit
View
Favorites
Tools
Help
SKIWeb (U) Predicted Conjunction Between METEOR ...
Page
Tools
SKiWeb
UNCLASSIFIED
Home
Actions
Event Logs
Communities of Interest
Force/Systems Status
Links
Help
Logout
Search
SKIWeb
for
Go
Advanced
Add Blog
Edit Event
Delete Event
Copy to New Event
Copy to New Related Event
Relate This Event to Another Event
Add To Critical List(s)
Add to My Event Log
Email Event
Read By
blog
UNCLASSIFIED
Internet
100%

Predicted Conjunction Between METEOR 1-20 vs THORAD AGENA D DEB 22/0404Z Mar 09 (Miss Distance is Now 0.105 KM)

220404Z Mar 2009 **GAPS SOCRATES**
(U) There is a predicted Conjunction between METEOR 1-20 and THORAD AGENA D DEB on 22/0404Z Mar 09 with a minimum range of 0.105.

**Impact:** 🟨 Moderate Impact to Mission Success

**Links:** Conjunction Details

**Keywords:** Conjunction , Satellite , GAPS , ProLogic

200501Z Mar 2009 **GAPS SOCRATES**
(U) New information update from SOCRATES:
Time of closest approach: 22/0404Z Mar 09
Closest Range: 0.112 KM
http://celestrak.com/SOCRATES/top10minrange.asp

190501Z Mar 2009 **GAPS SOCRATES**
(U) New information update from SOCRATES:
Time of closest approach: 22/0404Z Mar 09
Closest Range: 0.048 KM
http://celestrak.com/SOCRATES/top10minrange.asp

180501Z Mar 2009 **GAPS SOCRATES**
(U) New information update from SOCRATES:
Time of closest approach: 22/0404Z Mar 09
Closest Range: 0.093 KM
http://celestrak.com/SOCRATES/top10minrange.asp

Figure 1 A SKIWeb Event

Depending on the user's role in the organization, he or she will be interested in different subsets of events published on SKIWeb. The other main object class of SKIWeb is the *event log*, which is a list of events. The default view for users of SKIWeb is a reverse chronological list of all active events.

One of the philosophies behind SKIWeb was the desire for a flat information sharing network [Thaden 2006]. USSTRATCOM wanted information to flow freely without the confining influence of the chain of command. While flattening the organization from the information flow standpoint removed barriers to communication, it also removed an imposed social network that could help focus a user's attention on information that was relevant to his or her needs.

## PAL

The DARPA-sponsored Personalized Assistant that Learns (PAL) program brought together leading researchers in artificial intelligence, machine 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.

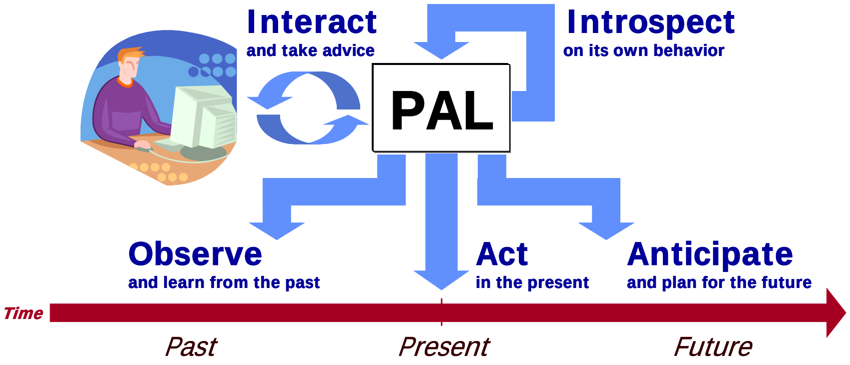


Figure 2 PAL Goals

The first three years of the PAL program focused almost entirely on cognitive assistance in an office environment. 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 the third year of this five year program, in parallel with continued research, concepts for the military application of the PAL technologies were formulated and a short movie was produced to stimulate thought by potential end-users. During the fourth year, it was decided between USSTRATCOM and DARPA, that the program would investigate if the technologies from PAL could enhance use of SKIWeb within its flat information sharing structure. The result was SKIPAL.

## Key SKIPAL Capabilities

### SKIPAL Learns

All of the capabilities below work because SKIPAL learns. SKIPAL currently uses three learning technologies from the PAL program: a recommendation engine, a topic modeler, and a text classifier. Each of these depend on instrumentation inserted into the system to provide data about the user’s behavior. SKIPAL tracks what events a user reads, authors, blogs to, and bookmarks. PAL also learns from requests to forward an event to another user, and from explicit feedback about the relevance of an event. Using the question and answer (Q&A) capability discussed below, SKIPAL also learns to model the relative expertise of users for topics. SKIPAL users can also train the system to recognize particular topics.

#### My Event Decision Learner (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. [Deans+ 2009]

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 a *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.

MEDL uses the training to generate a score between zero and one for each new event that represents a predicted relevance, where a higher score indicates a prediction that the user will be more likely to find the event relevant.

## 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 to have expertise on the event topic. The similar events and contributors are part of the SKIPAL enhanced event page described later. SKIPAL also uses this information to suggest users who might be able to answer a question that is asked on an event. SKIPAL uses the *iLink* topic modeler from the PAL program.

iLink [Davitz+ 2007] 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, or reading, an event. The weight of these signals can be adjusted to model different domains or use cases.

In SKIPAL, iLink has been used in several areas: 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.

## Text Classification

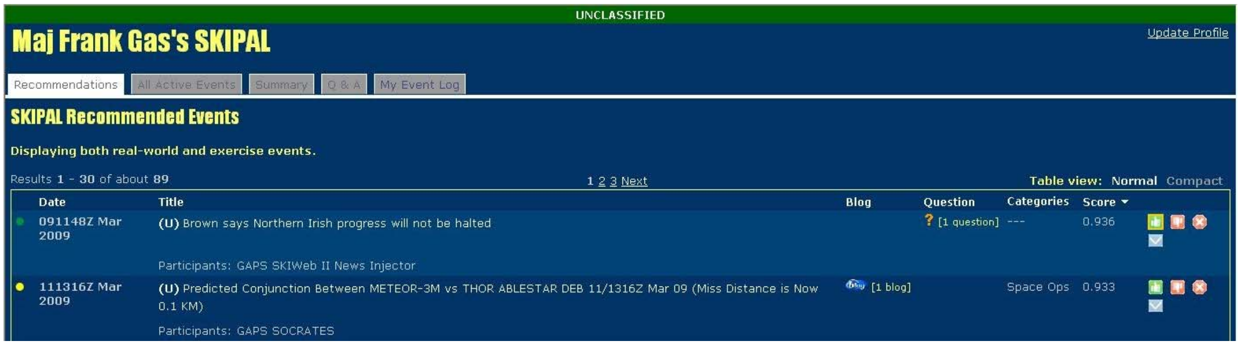
The Multi-Class Event Category (MEC) Classifier will classify events into multiple categories based on prior training by users. The categories represent common areas of interest to the SKIPAL community and are the same for each user. As such, every user benefits when MEC is trained. MEC is a multi-class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

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.

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."

## SKIPAL Recommends

The first user-facing capability provided to SKIWeb users was an event log based on event relevance rather than merely showing all active events. The figure below shows a list of events as recommended to a user by SKIPAL. Users interact with SKIPAL through a standard browser.



Maj Frank Gas's SKIPAL

SKIPAL Recommended Events

Displaying both real-world and exercise events.

Results 1 - 30 of about 89

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

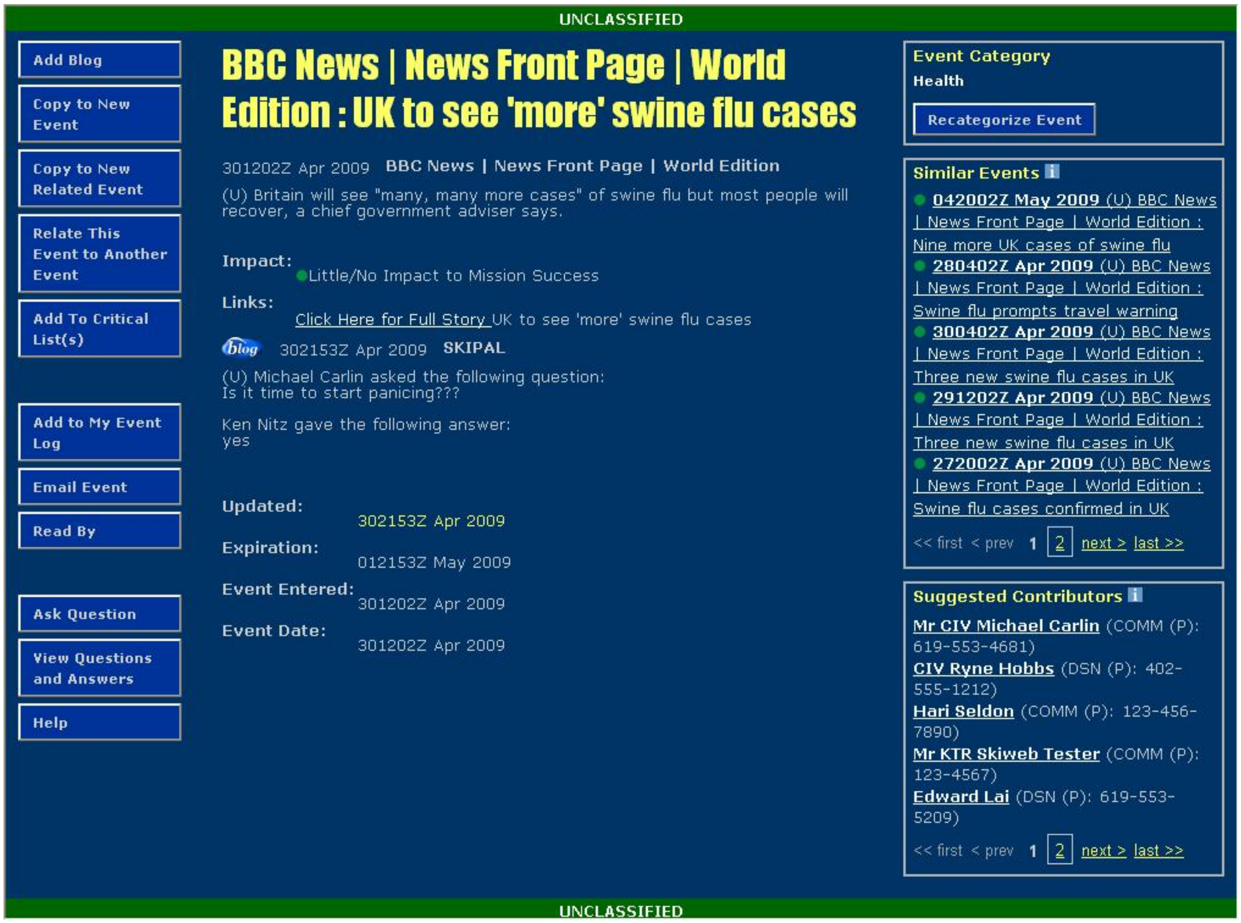
Figure 3 SKIPAL Recommendations

The Score column displays the predicted probability that the event is relevant to the user as computed by the MEDL recommendation engine. 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 an enhanced event page for that event (described later). By default, events on the recommendations page are sorted by decreasing MEDL score but can be resorted by clicking on any of the other column headings, except the Categories column.

Users 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.

## An Enhanced Event Page

The SKIPAL Enhanced Event page (an example is shown in the figure below) 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.



**BBC News | News Front Page | World Edition : UK to see 'more' swine flu cases**

301202Z Apr 2009 BBC News | News Front Page | World Edition

(U) Britain will see "many, many more cases" of swine flu but most people will recover, a chief government adviser says.

Impact: ● Little/No Impact to Mission Success

Links: Click Here for Full Story UK to see 'more' swine flu cases

blog 302153Z Apr 2009 SKIPAL

(U) Michael Carlin asked the following question:
Is it time to start panicing???

Ken Nitz gave the following answer:
yes

Updated: 302153Z Apr 2009

Expiration: 012153Z May 2009

Event Entered: 301202Z Apr 2009

Event Date: 301202Z Apr 2009

Event Category
Health

Similar Events

- 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

<< first < prev 1 2 next > last >>

Suggested Contributors

Mr CIV Michael Carlin (COMM (P): 619-553-4681)
CIV Ryne Hobbs (DSN (P): 402-555-1212)
Hari Seldon (COMM (P): 123-456-7890)
Mr KTR Skiweb Tester (COMM (P): 123-4567)
Edward Lai (DSN (P): 619-553-5209)

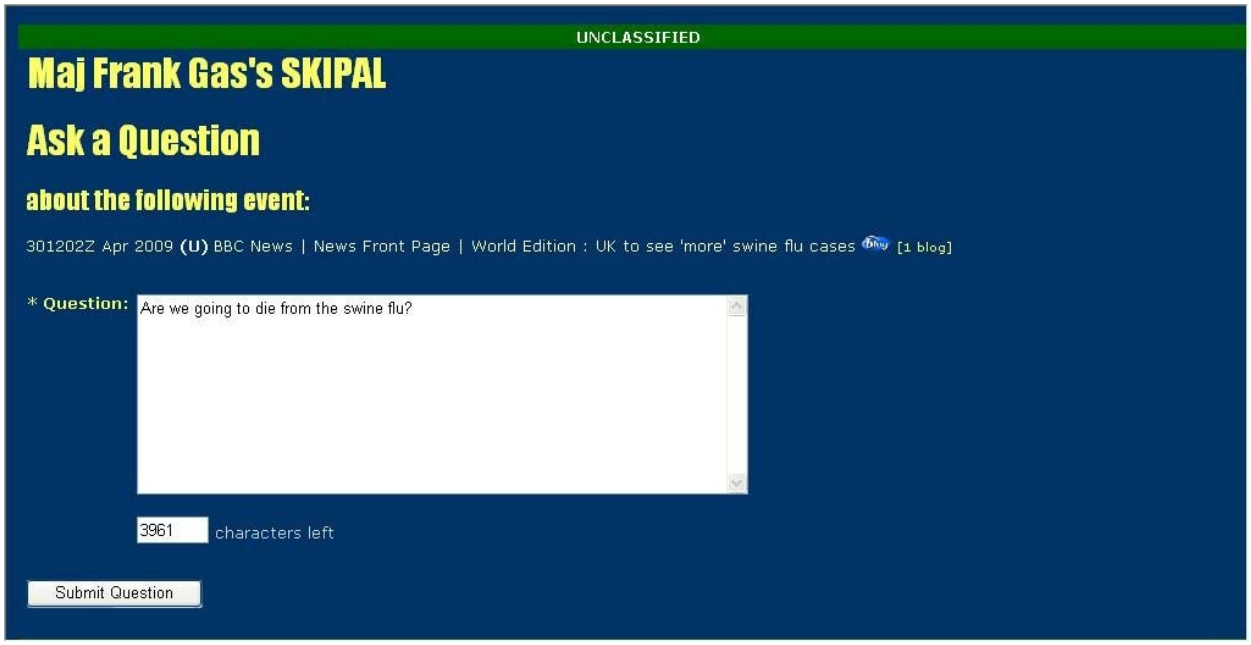
<< first < prev 1 2 next > last >>

Figure 4 Enhanced Event Page

The event's categories are shown in the upper right-hand corner. Clicking on the Recategorize Event button brings up a dialog window that allows the user to recategorize the event, further teaching the text classifier.

## Questions and Answers

The Ask Question button on the lower left of the enhanced event page is another enhancement to the basic SKIWeb capabilities. Clicking on the Ask Question button brings up the Ask a Question page in the figure below.



UNCLASSIFIED

Maj Frank Gas's SKIPAL

Ask a Question

about the following event:

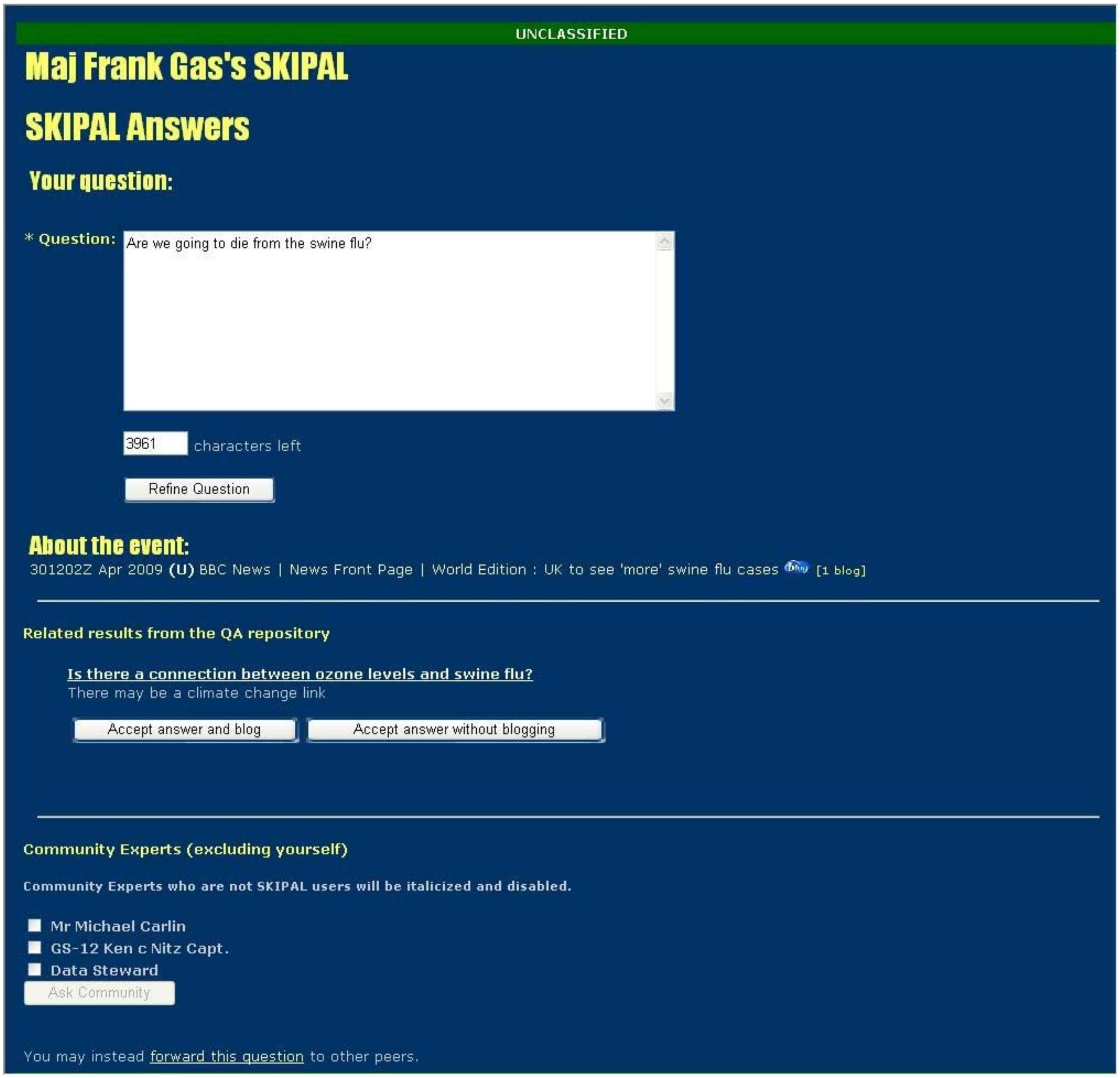
301202Z Apr 2009 (U) BBC News | News Front Page | World Edition : UK to see 'more' swine flu cases [1 blog]

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

3961 characters left

Figure 5 Ask a Question

When a question is submitted, SKIPAL provides similar questions that have been answered (shown below) to the user.


UNCLASSIFIED

Maj Frank Gas's SKIPAL

SKIPAL Answers

Your question:

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

3961 characters left

About the event:

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

Related results from the QA repository

Is there a connection between ozone levels and swine flu?
There may be a climate change link

Community Experts (excluding yourself)

Community Experts who are not SKIPAL users will be italicized and disabled.

□ Mr Michael Carlin
□ GS-12 Ken c Nitz Capt.
□ Data Steward

You may instead forward this question to other peers.

Figure 6 Questions and Answers

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. The listed experts are identified by the expertise model that is part of the topic model created by iLink. Additional pages allow users to track questions they have asked, and questions directed to them to answer. Users may forward questions they receive to others they believe have expertise in the topic. iLink tracks the forwarding actions as well and gives credit to those who know the person who came up with the satisfactory answer, boosting their expertise score for the topic.

## Architecture

SKIPAL is a standard J2EE web application. 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.

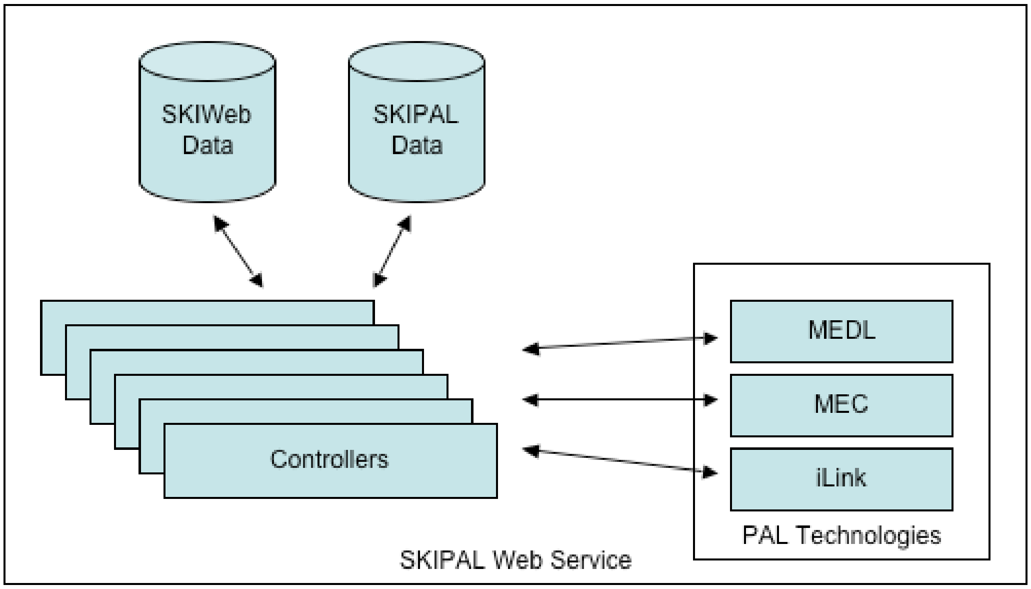


Figure 7 SKIPAL Architecture

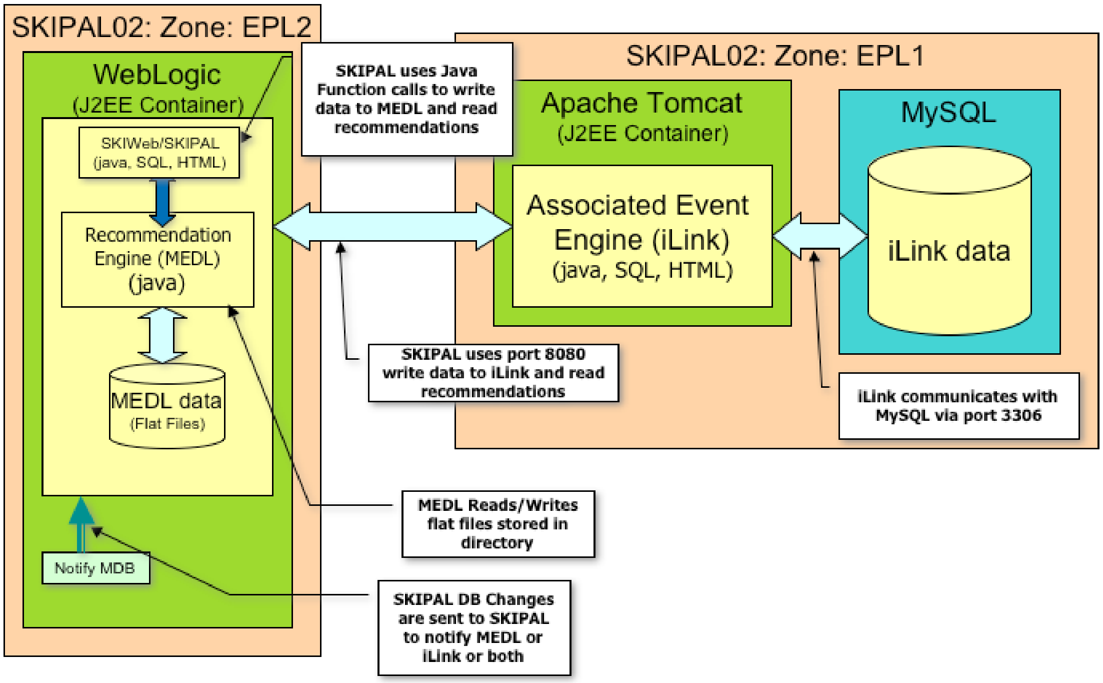

SKIPAL02: Zone: EPL2
WebLogic
(J2EE Container)
SKIWeb/SKIPAL
(java, SQL, HTML)
Recommendation Engine (MEDL)
(java)
MEDL data
(Flat Files)
Notify MDB
SKIPAL uses Java Function calls to write data to MEDL and read recommendations
SKIPAL02: Zone: EPL1
Apache Tomcat
(J2EE Container)
Associated Event Engine (iLink)
(java, SQL, HTML)
MySQL
iLink data
SKIPAL uses port 8080 write data to iLink and read recommendations
iLink communicates with MySQL via port 3306
MEDL Reads/Writes flat files stored in directory
SKIPAL DB Changes are sent to SKIPAL to notify MEDL or iLink or both

Figure 8 SKIPAL Component-level Design

SKIPAL operates with SKIWeb within the architecture shown below. The approach was developed to minimize impact on SKIWeb which is an operational capability in daily use over SIPRNET. An instance of SKIPAL is isolated in the Experimental Planning Laboratory (EPL) at USSTRATCOM. The addition of a firewall allows a controlled set of users to access the SKIPAL server.

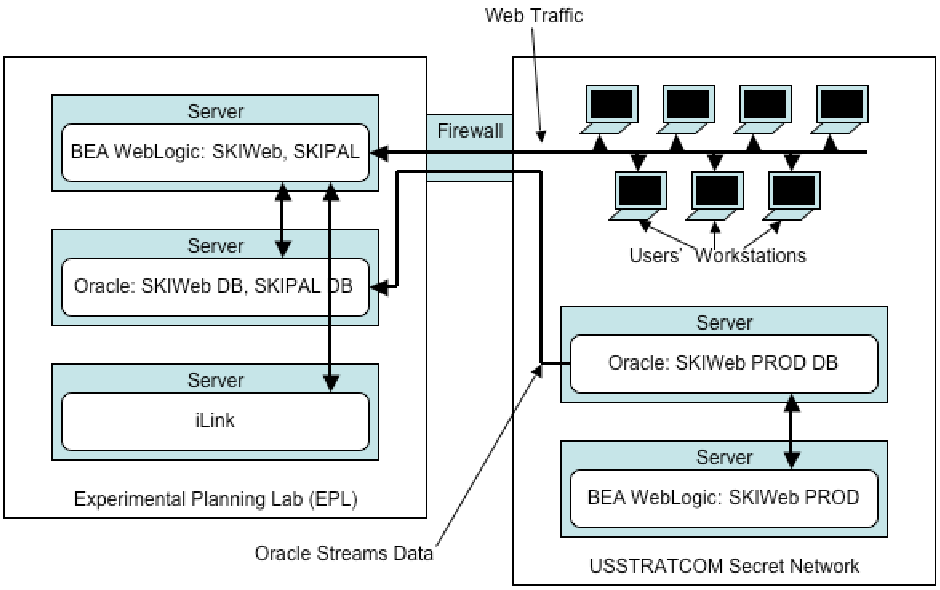

Web Traffic
Server
BEA WebLogic: SKIWeb, SKIPAL
Firewall
Server
Oracle: SKIWeb DB, SKIPAL DB
Users' Workstations
Server
iLink
Server
Oracle: SKIWeb PROD DB
Server
BEA WebLogic: SKIWeb PROD
Experimental Planning Lab (EPL)
Oracle Streams Data
USSTRATCOM Secret Network

Figure 9 SKIPAL-SKIWeb Deployment Architecture

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.

# Experimentation

## Spiral 2.1: Selecting the Recommendation Engine

### Experiment Design and General Results

In April 2008, the SKIPAL team 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, new data from the production SKIWeb was 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. The list was then displayed to the users on the SKIPAL Recommends These Events page.

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.

The figure below shows the average Receiver Operating 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. At this stage, the recommendation engine (EPL Recommended) in SKIPAL was based on iLink. It performed better than using the reverse chronological event log, but iLink technology was not a good fit for recommending events. Further tests on the same data using recommendation engines based on other technologies provided evidence used to select MEDL as the basis for our recommendation engine (WordCount line in graph).

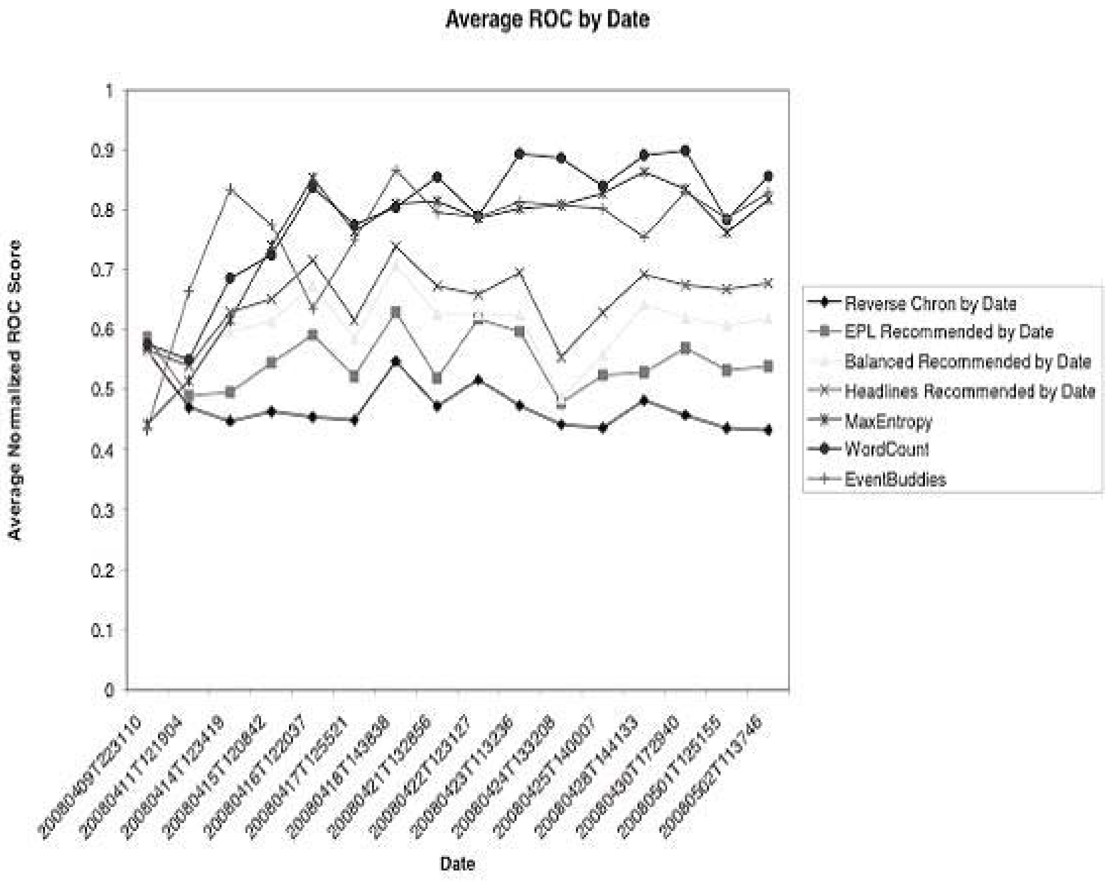


Figure 10 Average ROC Scores for each recommendation engine tested.

## Detailed Technical Approach

To analyze the data, we first 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 inTable 1. Note that different thresholds result in different confusion matrices (and, hence, different classifiers).

Table 1 Confusion matrix for a given threshold

| | Relevant to the user | Not relevant to the user |
|---|---|---|
| Classified as relevant | *a* = # of *TP*s | *b* = # of *FP*s |
| Classified as not relevant | *c* = # of *FN*s | *d* = # of *TN*s |

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 the table above. 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. All of the individual ROC, AUC, precision, and recall curves are documented in [Deans+ 2009].

## Further Examination of the Recommendation Engine

### Experiment Design for Spiral 2.3

Moving to live day-to-day operations provided a challenge with regard to evaluating the recommendation engine. In the April 2008 experiments, we could ask the users to review each and every event and mark it relevant or not relevant. This could not be done with a larger group of users or for a longer test where the system would not be stopped and loaded each day. Measures of ROC and recall require knowledge of all the events at any given time.

To solve this problem, we created an approach to sample the data. We added a survey page that test users would be frequently shown. 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.

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 2008 –February 2009 (Spiral 2.4) experiments. During the August–October (Spiral 2.3) 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. The figure below 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.

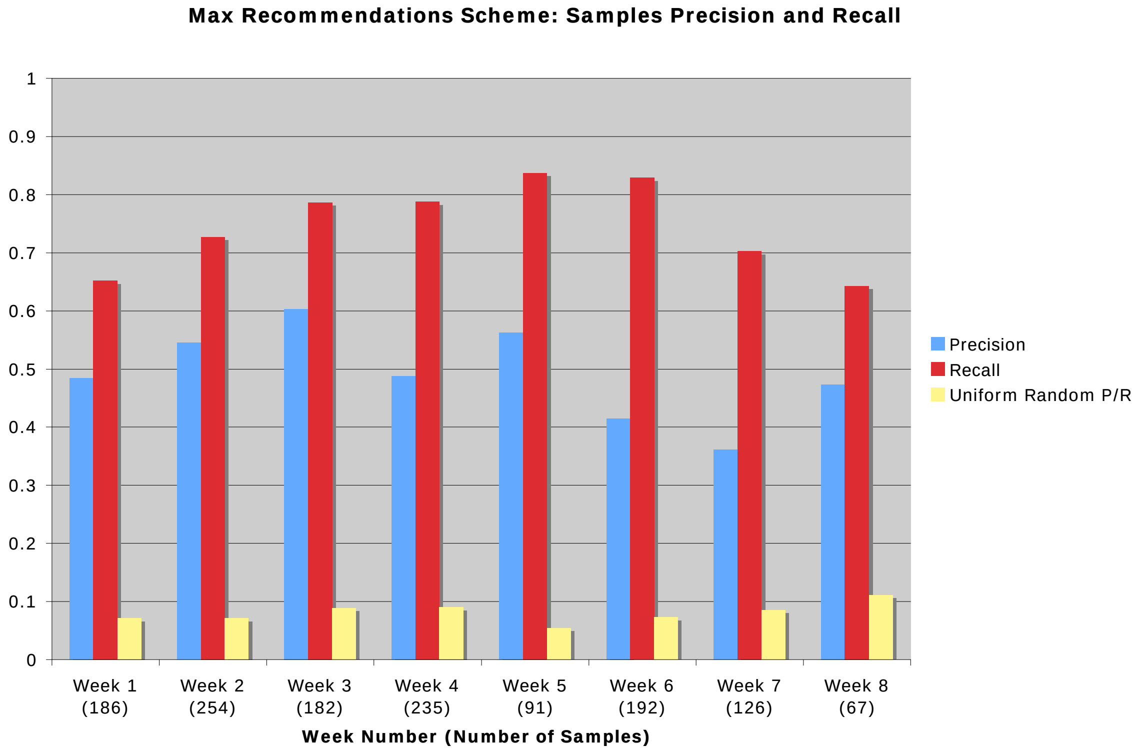


Figure 11 Estimated Precision and Recall for Spiral 2.3

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.

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 11 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.

## Experiment Results for Spiral 2.4

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 the figure below. 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.

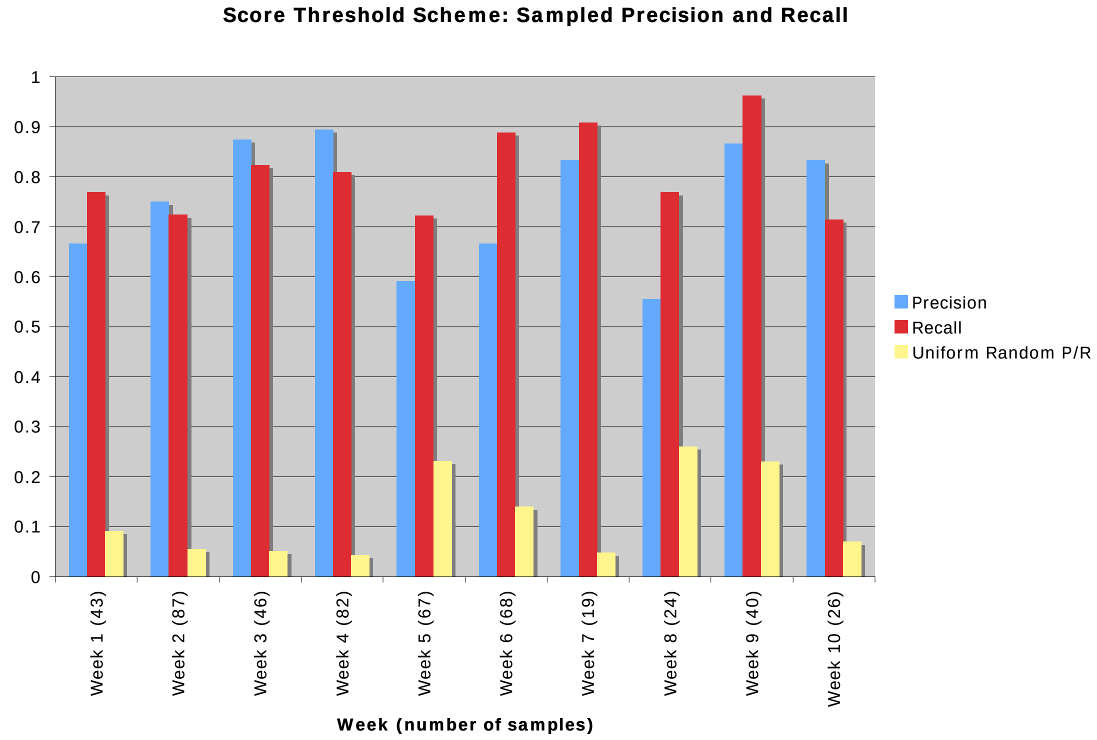


Figure 12 Estimated Precision and Recall with Score Threshold Scheme

The full detailed results for both spiral 2.3 and 2.4 experiments can be found in [Deans+ 2009].

## Conclusions

The data from our experiments as well as user feedback, provided ample evidence that SKIPAL is learning what information is relevant to each user and providing a useful service to the users. By helping users find relevant people to an event or question, SKIPAL exposes a social network that represents how information really is developed within the virtual organization represented by SIPRNET users of SKIWeb. This social network is far more relevant to information sharing than an imposed chain of command that has more relevance for task accountability. The recommendation engine has proven to help users focus on events that are related to their work and therefore improve efficiency in dealing with the large quantities of information competing for their attention.

SKIPAL was demonstrated to General Cartwright (Vice Chairman, Joint Chiefs of Staff) to show him what was started as a result of his requirements when he was Commander of USSTRATCOM. Soon after, USSTRATCOM submitted the required program request documentation to allow SKIPAL capabilities to be fully integrated into SKIWeb. This will provide the final transition step from DARPA research to fully supported operational capability.

# References

[Thaden 2006] Thaden, F., “Blogs v. Freedom of Speech: A Commander’s Primer Regarding First Amendment Rights As They Apply to the Blogosphere”, AU/ACSC/8234/AY06, Air Command and Staff College, Air University, April 2006, accessed at www.au.af.mil/au/awc/awcgate/acsc/thaden.pdf on 8 January 2010.

[Deans+ 2009] Deans, B., Keifer, K., Nitz, K., Tam, L., Carlin, M., Lai, E., Lange, D., Ling, A., Bolton, J., Graves, B., and Reestman B., “SKIPAL Phase 2: Final Technical Report”, Technical Report 1981, Space and Naval Warfare Systems Center Pacific, November 2009.

[Davitz+ 2007] Davitz, J., Yu, J., Basu, S., Gutelius, D., and Harris, A., “iLink: Search and routing in social networks”, *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, CA, August 2007.