GeoTracker: Geospatial and Temporal RSS Navigation

Yih-Farn Chen, Giuseppe Di Fabbrizio, David Gibbon, Rittwik Jana, Serban Jora,
Bernard Renger, Bin Wei
AT&T Labs - Research
180 Park Ave
Florham Park, NJ, 07932
+1 973 360 8653

{chen, pino, dcg, rjana, jora, renger, bw}@research.att.com

ABSTRACT

The Web is rapidly moving towards a platform for mass collaboration in content production and consumption. Fresh content on a variety of topics, people, and places is being created and made available on the Web at breathtaking speed. Navigating the content effectively not only requires techniques such as aggregating various RSS-enabled feeds, but it also demands a new browsing paradigm. In this paper, we present novel geospatial and temporal browsing techniques that provide users with the capability of aggregating and navigating RSS-enabled content in a timely, personalized and automatic manner. In particular, we describe a system called GeoTracker that utilizes both a geospatial representation and a temporal (chronological) presentation to help users spot the most relevant updates quickly. Within the context of this work, we provide a middleware engine that supports intelligent aggregation and dissemination of RSS feeds with personalization to desktops and mobile devices. We study the navigation capabilities of this system on two kinds of data sets, namely, 2006 World Cup soccer data collected over two months and breaking news items that occur every day. We also demonstrate that the application of such technologies to the video search results returned by YouTube and Google greatly enhances a user's ability in locating and browsing videos based on his or her geographical interests. Finally, we demonstrate that the location inference performance of GeoTracker compares well against machine learning techniques used in the natural language processing/information retrieval community. Despite its algorithm simplicity, it preserves high recall percentages.

Categories and Subject Descriptors

H.4.3 [**Communication Applications**]: Information Browsers.

General Terms

Algorithms, Measurement, Experimentation, Human Factors, Languages.

Keywords: RSS, geospatial tagging, blog, multimedia.

1. INTRODUCTION

The emergence of RSS (Really Simple Syndication) [1] technologies and web logs (a.k.a., blogs) have helped transform the Web into a service platform that allows normal Web users to compete with traditional news media for timely content publication, aggregation, and delivery. Web users are not satisfied with simply finding out about X through Web surfing but are also interested in finding out *what's new* about X, whether X is a person, a place, an event, a topic, or any entity.

Besides writing their own blogs, users also collaborate and make contributions to the Web content by tagging, *digging*, and other means to help label, rank, and prune Web content. Notable examples are Wikipedia [20], which allows collaborative authoring, Flickr [21], which is a photo sharing website, and Digg [10], which allows democratic editorial control. The result is massive collaboration in content production and consumption.

Increasingly, updates of these entities are provided through RSS feeds; unfortunately, as the number and variety of RSS feeds grow exponentially over time, finding relevant updates quickly requires a paradigm shift on how we browse Web information.

In this paper, we present our studies on aggregating and visualizing RSS feeds in a *geospatia*l and *temporal* manner, a major departure from the traditional text-based newspaper layout, which is still the dominant style on most news websites. Our contributions are specifically

- Presenting RSS data along both geospatial and temporal dimensions.
- Aggregating RSS feeds between structured, edited content and non-structured, typically informal blog sites.
- Creating Media RSS feeds with automatic extraction and inference of location coordinates.
- Mining RSS data to provide personalized search capabilities and extracting salient themes.
- Formulating a framework to compare the performance of GeoTracker, which uses a simple rule-based classifier, with more sophisticated classifiers and to motivate a case for the former with respect to RSS data.

Our approach is to build a middleware platform, MxM that integrates these technologies to provide a geospatial and temporal RSS browser that enables users to quickly browse updates in a personalized manner from a variety of devices. Section 2 explores this new content navigation paradigm in detail. Section 3 describes the GeoTracker system architecture based on the MxM platform and the integration with the MIRACLE multimedia content management platform. Section 4 describes the experiments and performance evaluations on GeoTracker, followed by discussions on related work, future work, and conclusions.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

2. CONTENT NAVIGATION

Navigating RSS content is typically performed today using RSS-aware programs called *readers*. They are usually available as stand-alone programs or extensions to web browsers (plugins). The readers fetch recent updates periodically from a list of user-subscribed sites and alert the users accordingly. A user is typically notified of recent updates by means of a popup window showing a short textual description of the RSS item.

In this section, we introduce a new way of presenting RSS data, with a geospatial nature (i.e., over locations on a world map at a given point in time) and a temporal nature (i.e., over time at a particular location or region). We believe the idea of geolocating RSS items as described in this paper is novel. By geolocating RSS, we mean associating RSS items on a world map where we perform reverse mapping of each story to multiple locations. Similar ideas in the context of books have been looked at for locating places mentioned in books [5].

The RSS and blogging community have added the Geotagging (sometimes referred to as Geocoding) extension to RSS, which adds explicit geographic coordinates (latitude, longitude) to each feed entry. A proxy approach can also annotate a regular RSS feed with geotags by mining the text. Flickr encourages users to use geotags and there are websites that combine Google Maps and geo-tagged feeds such as flickrmap [21]. It appears that in most of these works, each RSS item is mapped to only one relevant location, but a news story concerning the Israel-Lebanon conflict should be highlighted for both Israel and Lebanon. We take the approach that GeoTracker must handle *many-to-many* relationships (in the notion of the Entity-Relationship model) between locations and news items.

2.1 GeoTracker: Geolocating RSS

Most users browse news items by starting from a news website that lists the latest news in the order of importance and freshness. Our new browsing approach presents a paradigm shift in that we display RSS data in a geographic presentation layer. The browser allows the user to navigate (zoom, pan) the RSS view on a world map. The geo-mapping software utilizes the Google mapping service [28] to render locations on the map automatically. Further details on extracting and inferring location information are provided in the following subsections.

2.1.1 Mining location information from text

Our approach here is based on the assumption that the location information is explicitly expressed in the text and it is most easily inferred from proper names such as France, U.N., or New York. We implemented a rule-based tagger that extracts names from any given text. We consider a sentence in the text as a set of words and a proper name is defined as a set of capitalized words separated by certain predefined characters.

The names resulting from the tagger are first matched against a location database (such as an XML representation of the 'Mondial' database [19]) that lists all countries, provinces for each country, and the most important cities for each country. Cross-referenced information is also available: capitals (country, city), etc. Alternative names are accepted, for example:

USA, United States, U.S.

Havana, La Habana

Munchen, Munich

The matching process looks for exact, case-insensitive matches, which are then translated into locations. During this step we perform some cleaning and reduction operations. For example, we eliminate generic matches in favor of more specific ones: a match for 'France' will be eliminated in favor of a match for 'Lyon', a particular city in France (this type of reduction is done with respect to matches resulting from a single RSS item). Finally, locations are reduced to their canonical form *(city, country)*, (e.g., 'New Jersey' would be reduced to (Trenton, USA)).

Application-specific knowledge can also be added as separate modules to match names to locations. A 'world politics' module can match state official names (presidents, prime ministers, etc.) and organizations (U.N., WTO, etc.) to their locations; a sports module can match player and coach names to the locations of the clubs and countries for which they are playing. In this domain-specific module, the locations must be specified in the canonical form (city, country). This enhancement can improve location inference precision, but our observations show (see Section 4.4.1) that even without these additional modules, GeoTracker is still very accurate.

2.1.2 Presenting RSS feeds on a world map

With a user profile that stores the user's interests, we can personalize a user's browsing or viewing experience by highlighting events and presenting associated multimedia items that match his or her geographic interests. For example, Figure 1 shows the locations that correspond to Yahoo top news items. Yahoo generates RSS feeds segmented by categories, one of which corresponds to breaking news items. The interface allows users to zoom in and out of any area and to pan from one side to the other easily. The zoom and pan features are capabilities that we inherit from Google Maps. Clicking on the pin on Mexico leads a user to the recent Yahoo news page on Mexico, as shown in Figure 1. Note that if multiple locations are mentioned in a story, GeoTracker will display multiple pins that correspond to the same story on the world map.

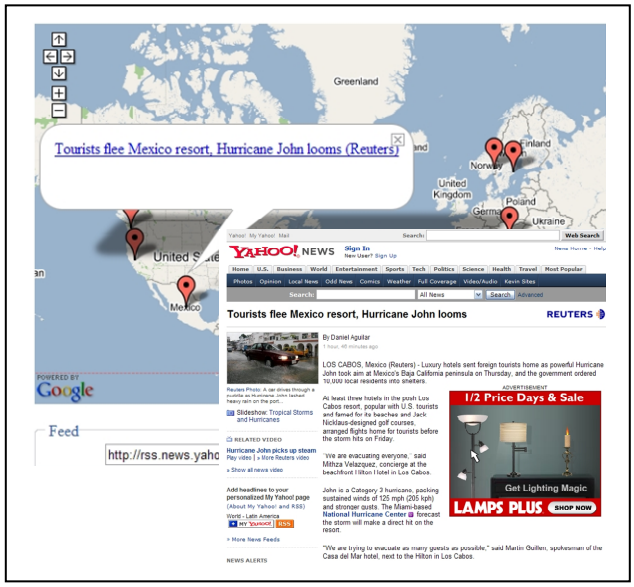

Figure 1 - Geolocation of international top stories from Yahoo RSS feeds

2.1.3 Browsing Latest Multimedia Content

As we can see from Figure 1 there are frequently video clips associated with news items. In addition, video search sites like

Google Video and YouTube allow one to find video clips related to any search term and provide video feeds following the MediaRSS specification of Yahoo [18]. Using the same user interface, GeoTracker can take these RSS feeds and provide pins that map to the actual video clips, audio clips, or other multimedia content. The information bubble can contain links that offer a transcoded version of the multimedia clip through the MIRACLE platform (see Section 3.2). A user may choose the resolution appropriate for his or her device. Figure 2 shows the recent videos available on the Google website that are related to cities or towns in India.

Figure 2 - Geolocation of Google videos published in India

Each location found in the feed is a pin on a Google Maps area. The information bubble that opens at each pin shows the items relevant to that location. In order to present the *m-by-n* relationship between events (feed items) and locations, we added a Google Maps picture-in-picture (PIP) 'highlights' feature that, when an event (NBA matches in this example) is selected at a particular location's information bubble, the PIP displays all locations that correspond to that story. Figure 3 illustrates this feature for 2 NBA games showing four team locations.

2.2 GeoTracker: Visualizing RSS over Time

The approach described in Section 2.1 allows users to view the geographical distributions of events during any particular snapshot (or a short period of time). It does not show how the events evolve over a period of time.

We have developed another visualization approach where a sliding bar of time (say over a time range of one day or one week) is added to the browser. As the user moves the slider on the time scale, the geographic distribution of top news events may change over time. In addition, as more events occur at the same location, the color coding of that location (or the pin associated with it) changes to reflect the intensity of news events happening there. Figure 4 shows the browser interface for GeoTracker with the slider – more examples are presented in Section 4.

3. SYSTEM ARCHITECTURE

This section details the implementation of GeoTracker. We start with a middleware platform, Mobile Multimedia Middleware (MxM), which provides the foundation to implement the GeoTracker functionalities. A content management platform, MIRACLE [12][13], is also used in conjunction with MxM to perform MediaRSS [18] enabled content search and retrieval. The architecture is designed to be flexible enough to incorporate other types of media and is not limited to only RSS content.

Figure 3 - Picture–in-Picture showing events at multiple locations

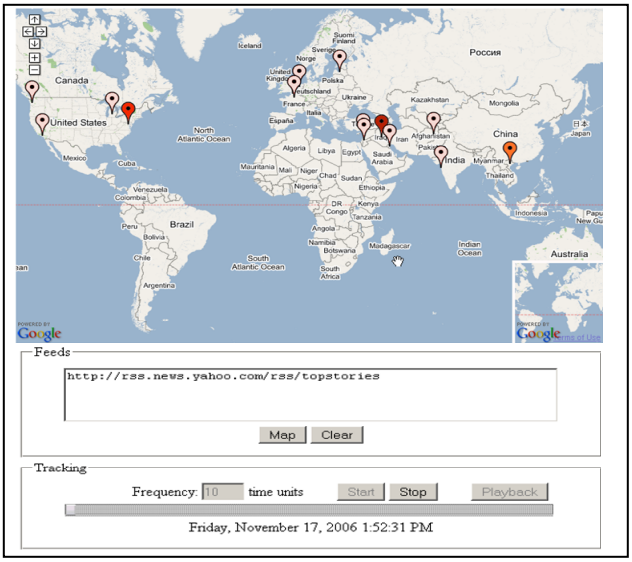

Figure 4 - GeoTracker: Temporal mapping with time slider

3.1 The MxM Middleware Platform

As access to the Web is migrating from a desktop-only environment to a mix of various computers and mobile devices anywhere, we must allow easy submission of blogs and tracking requests from mobile devices, and easy dissemination of aggregated and filtered content to devices based on their capabilities. We have designed a mobile multimedia content aggregation and dissemination platform, MxM, to support such ubiquitous publication and retrieval of blogs and RSS content.

The MxM platform (see Figure 5) is a middleware system that contains gateways, servers, a message switch and databases. Gateways send and receive messages or data to and from devices using different protocols (e.g., http, mail, sms, mms, voice, fax, SIP, instant messaging, etc.). Messages or requests received from these gateways are authenticated to identify, whenever possible, the sender, the user agent, and device profile [26], and then transmitted through the message switch to any of the MxM servers. Each server hosts an identical set of *infolets* that

implement specific application logic and usually provide access to one or more RSS or non-RSS information sources. An infolet's output needs to conform to the destination delivery context for a session established for the user's device. The content can also be personalized based on the user profile stored in the profile database and adapted according to the device profile.

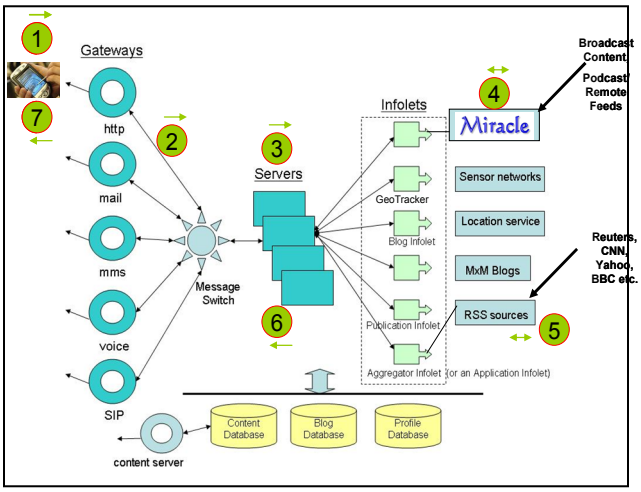

Figure 5 - MxM platform with the GeoTracker infolet

The MxM platform offers support for information transcoding (format conversion) in the form of a framework that can be used by the infolet provider. For example, the blog infolet converts a blog entry submitted by a user through any of the gateways into a blog information item stored in the Blog database. RSS-enabled infolets implement protocol interfaces that access various information sources (such as a location service, a RSS information source, RFID sensor service, etc.). Such infolets also include and implement a module that converts the retrieved information from the various services into a properly formatted RSS data feed. In general, the conversion of retrieved information into a format that facilitates creating RSS feeds is performed in a timely manner by an infolet making the data amenable for aggregation with other sources of information. The gateways which have interfaces to support various devices can handle the content dissemination to users, either with a desktop or a mobile device. Additional details about the predecessor of the MxM platform (without RSS integration) are described elsewhere [4][29].

3.2 The MIRACLE Content Management Engine

For content queries, MIRACLE supports OpenSearch which was developed by A9 [14] and defines a common API for applications such as GeoTracker to use when querying several different search engines. OpenSearch requires that the results be formatted in RSS and namespace extensions are supported. For MIRACLE, MediaRSS is the logical choice for the greatest flexibility with video retrieval applications; however we also support the iTunes namespace extensions due to the popularity of Podcasts and Apple's iPod. When MIRACLE ingests MPEG-2 DVR Broadcast TV content at 6Mb/s, we form five different representations of the media at different bandwidths ranging from 2Mb/s down to 64Kb/s for various applications. The lower bitrate streams contain only the demultiplexed and transcoded audio and are intended for devices that cannot play video, such as the Apple iPod Shuffle, or for applications where audio playback is more appropriate for the user interface than launching a video player. The MediaRSS [18] `media:group` element with its `media:content` subelements is ideal for representing these multiple encodings of a particular media asset.

MIRACLE users can create their own RSS feeds tailored to their personal preferences by specifying query terms using an intuitive query syntax. The queries can be as simple as a single word or phrase, or may include additional content metadata restrictions such as specifying a date range or a set of television programs to search. If desired, queries can also include negation and grouping of directives. This framework can be extended to include geospatial restrictions in the metadata as well.

In addition to generating RSS search results, MIRACLE uses RSS for content ingest. Podcasts are the richest source of RSS-described syndicated content on the Web today. MIRACLE maintains a list of Podcast URLs, associated content owner information, and a program naming schema for recurring scheduled broadcast content. On a scheduled basis, each of the feeds in the list is checked for new content. If any is found, the content is downloaded, processed and entered into the media archive.

Given that MIRACLE can ingest RSS feeds and perform content-based indexing and retrieval, and that the output can be formatted as an RSS feed as well, one can think of MIRACLE as an RSS filter/aggregator. Users can specify topics of interest and this information is stored in a profile, and then the system returns only RSS feeds that are related to their interest. This function is common in RSS readers, but usually the filtering is based only on RSS metadata. For textual (HTML) RSS feeds, standard information retrieval methods can be used. MIRACLE extends these concepts to apply to media feeds.

3.3 MxM-MIRACLE interactions

A typical user may navigate Web content from his desktop or a mobile device. MxM captures the user search term and presents this information to MIRACLE for querying. Results particular to the search term are returned to MxM via a MediaRSS feed. MxM aggregates this information with other RSS feeds that are ingested into the platform. The collective feeds are then processed to remove duplicates and then the GeoTracker infolet extracts and infers locations as per Section 2.1. A world map is then presented to the user with a summary of all results relevant to this search query. A typical use case scenario is shown in Figure 5:

1. A GeoTracker request (keywords) is submitted by a user from a mobile device.
2. The HTTP gateway in MxM forwards the request to one of the servers through the message switch.
3. The server performs the necessary processing to invoke the target GeoTracker infolet, which in turn invokes other infolets (Miracle infolet and Aggregator infolet) to finish the task.
4. Query to MIRACLE for media content that matches the request.
5. RSS aggregator feed is used to extract additional relevant information

6. The GeoTracker infolet extracts and infers location information from collective feeds obtained from MIRACLE and external RSS sources.
7. Results are transcoded and presented on a map and sent back to the end user.

3.4 Data Collection

To test the effectiveness of GeoTracker on real-life events, we acquired the following data:

1. Video recordings of all 64 2006 World Cup Soccer matches, including pre- and post-match programs.
2. Top 100 World Cup Soccer related RSS feeds and blogs around the time of the tournament.

The World Cup Soccer matches were recorded and processed by the MIRACLE system [13]. The MIRACLE system uses automated content-based media processing algorithms and systems to collect, organize, index, and repurpose video and multimedia information. The MIRACLE user interface contains many useful features that allow the user to search for relevant information in the World Cup matches. Search retrieval results can appear as Web pages or as RSS 2.0 feeds with Media RSS extensions. All 64 World Cup Soccer matches were recorded and the average length of a match was around 2.5 hours.

World Cup Soccer RSS feeds were collected for 35 days from over 100 sources with a refresh rate of one hour. The sources included the official World Cup news outlets, different blogs, Google news, etc. Some feed sources were organized by participating countries while others were generic sports feeds.

3.5 Data Processing

Upon a client request, data acquisition and processing occurs within the GeoTracker infolet and associated infolets where RSS feeds are reduced to a unique set and annotated with location information before being presented to the end user.

3.5.1 RSS Data Processing/Geocoding

The main functionality of GeoTracker is mining the aggregated RSS feed for location information and annotating the feed with the extracted information. The text mining, location extraction, and other aggregation-related aspects (eliminating duplicate items) are done 'on the fly' as the RSS feeds are passed through XSLT transformations that are MediaRSS-aware. The resulting RSS feed keeps only the elements necessary to render the map. GeoTracker, as an MxM infolet, takes into account device, user, and service profiles and adapts the content accordingly for the target device and user.

3.5.1.1 Annotating the RSS feed

GeoTracker maps each RSS feed item to a set of locations and vice-versa: each location can be referenced by a set of RSS items in the feed (an *m-to-n* relationship between feed items and locations).

Each reference to elements in the locations index is added to the end of the resulting RSS document. As every RSS item can mention several locations, multiple such location references can be added to the item's locations set. The locations index contains all the locations identified within the aggregated feeds. A location element found in the index contains the canonical location name (city, country) and the geocoding information (if available, geocoding is added on the server side for known locations). The `refs` attribute serves as a reference counter and indicates how many distinct feed items were referring to that location (see Figure 6).

```
<rss><channel>
<item>
 <title>Concert goer throws drink at Streisand (AP)</title>
 <mrss:content height="105"width="130" type="image/jpeg" url="" />
 <mxm:locations><mxm:location idref="8"/>
                <mxm:location idref="9"/>
</mxm:locations>
</item><item>
 <title>Mandela leads tributes to S.Africa's Botha (Reuters)</title>
 <mrss:content height="99" width="130" type="image/jpeg" url="..." />
<mxm:locations>
<mxm:location idref="15"/> </mxm:locations>
</item><item>
 <title>Mandela praises arch enemy Botha after apartheid leader dies (AFP)
</title>
 <mrss:content height="130" width="104" type="image/jpeg" url="..." />
 <mxm:locations>  <mxm:location idref="15"/> </mxm:locations>
</item>
<mxm:locations>
   <mxm:location id="8" refs="1">Fort Lauderdale,U.S.</mxm:location>
   <mxm:location id="9" refs="1">Toronto,Canada</mxm:location>
   <mxm:location id="15" refs="2">Pretoria,South Africa
    <mxm:geo>  <mxm:lat>-25.75</mxm:lat>
               <mxm:lng>28.2333</mxm:lng></mxm:geo>
   </mxm:location>
 </mxm:locations>
</channel></rss>
```

Figure 6 - GeoTracker infolet MediaRSS output for client presentation

3.5.2 Presentation

The client side deals with presentation aspects and is based on the JavaScript and Google Maps API. After a set of RSS feeds is submitted to the server, AJAX style, the resulting geo-annotated aggregated feed in Figure 7 is presented to the user. Those locations for which no geocode information was supplied by the MxM server are submitted to Google's geocoding service.

All geocoded locations are marked on the map with a pin and the associated pin-info window presents the events (RSS items) referencing that location. Client side processing is MediaRSS aware and will present the user with icons for each type of recognized content: image, audio, video. This allows direct access from GeoTracker to the multimedia content related to each item at a particular location. The tracking feature as described in Section 2.2 is implemented as a time driven set of basic geomappings where the presentation of the data accumulated in the background is driven by the slider control or through a continuous playback feature. Histograms are also supported as part of the tracking feature. It builds an SVG document reflecting the number of events at a given location over time, which is translated into a JPEG with the help of the Apache's Batik framework.

Presentation on less capable user agents (in particular, those not supported by Google Maps) can be achieved, with some compromises, by using MxM's infolet-based presentation and content transcoding capabilities.

3.5.3 Video Data Processing

The details of the MIRACLE processing are described elsewhere [12][13]. During the processing stage, various transcoded formats of the original audio and video are created so that the format appropriate for different devices/interfaces can be used. The metadata (show owner, title, description, etc.), Closed Caption (CC) text, as well as image data are also extracted.

4. EXPERIMENTS AND RESULTS

We have conducted several experiments to illustrate and evaluate some features of the platform with a focus on the spatial and temporal aspects.

4.1 GeoTracking a Corpus of Data: World Cup 2006

Figure 7 illustrates the capability of the platform to search a corpus of collected data over a period of time and present an interesting aggregated RSS feed to the user. We show the query for "Klose", the top scoring World Cup 2006 soccer player. In this scenario, a user is interested in following a player, namely Miroslav Klose, and viewing his goals throughout the tournament. We annotate Figure 7 with labels that show RSS data mining results:

1. Feed title – this appears in the media RSS document that links to the search engine MIRACLE shown in Figure 8.
2. Video clip – This is a link to a video stream corresponding to the search term (in this case, the first occurrence of the goal scoring video clip).
3. Thumbnail – This is a link to an image snapshot of one of the goals of Klose (see Figure 8).
4. Audio clip – This is a link to the audio stream corresponding to the search term.
5. MediaRSS URL – This is a link to the media RSS URL that is shown on the map.

Figure 8 shows a page returned from the Miracle search engine that corresponds to the MediaRSS URL with the following annotations:

1. Feed title
2. Search term – This is the proximity search criteria used to show all instances of "Klose NEAR goal".
3. Program segment number (pagination) - This is a link to a particular hits page within the complete show.
4. Hits timeline – The dots represent links to segments (hits) that match the search term.
5. Thumbnail, video clip – This is a link to the media for a particular segment.
6. The RSS feed is derived from the MIRACLE search engine which is ingested by the MxM platform. For illustration purposes this URL is shown in Figure 7 (item 5).

4.2 U-Bar: GeoTracker and Blog Integration

Figure 9 illustrates the concept of integrated blogs and geomapping. Often, as information is presented on a news site, the real time dynamics on how the event and public opinions evolved over time is lost. With the introduction of RSS, a dynamic presentation can offer the realism of experiencing a breaking news event. This example shows the geomap of the famous head butting incident of Zidane. As this incident was unfolding, blog sites were capturing comments from across the world. There were formal edited blog sites like CNN and ESPN but there were also some interesting informal opinions from the public. The integration is illustrated in the form of a U-shaped bar where the blog opinions are aligned on the left of the geomap while the bottom shows the the expanded view of a particular blog item. Relevant images from various blogs are also shown to the right of the map.

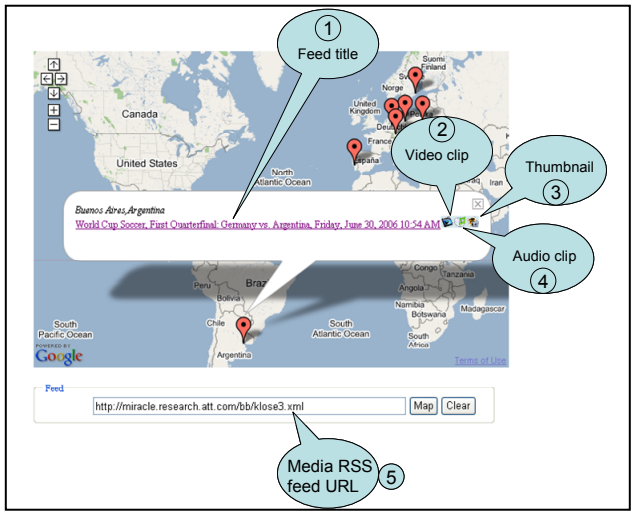

Figure 7 - RSS data mining and visualization from a corpus of collected 2006 FIFA World Cup data

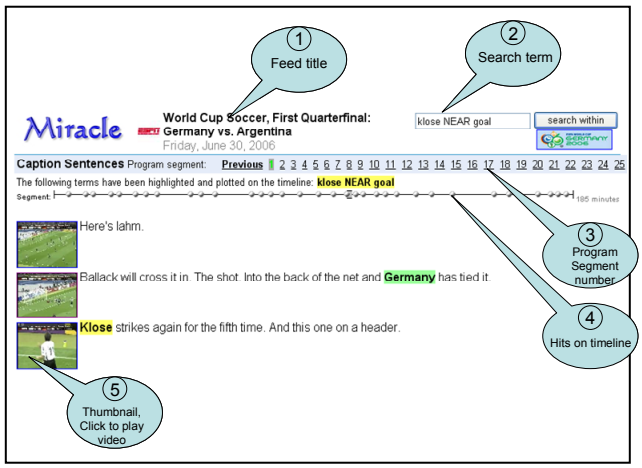

Figure 8 - Feed title link to MIRACLE search engine

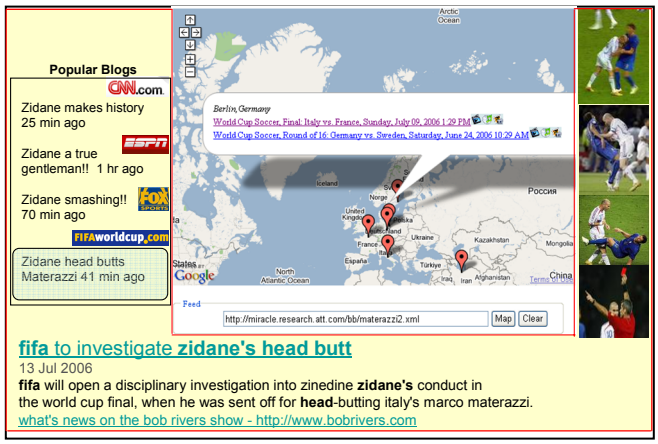

Figure 9 The U-Bar concept - integrated blogs and geomap

4.3 GeoTracker: Temporal Navigation

Figure 10 shows how we can track the reach and prominence of a particular event as it is played back in time or the geographical distribution of all top news events. The platform allows a user to record RSS updates and replay these items on a world map. Items

that are recent or appear in high frequency are highlighted with a blinking icon. The slider allows the user to also rewind or fast-forward to a particular point in time. Figure 10 shows two snapshots taken over the period of one particular day in November 2006. The user can simply move the slider bar to browse the event distribution during different points in time.

Figure 10 - Time Slider: Tracking evolution of events over time with GeoTracker

4.3.1 Theme independent – Strength vs. time, fixed at a particular location (histogram).

Figure 11 shows a plot of the prominent events that evolve over time at a particular location, in this case Iraq. News broadcasters like CNN try to capture global headlines while the local media broadcasters concentrate on local events for a particular geographic domain. This is theme or subject independent. The advantage of geomapping events occurring at a particular location can be presented to emphasize the importance of breaking events local to a community, which is not always obvious, since the public tends to focus on global events. In this figure, we illustrate the capability to track events over time by means of a time navigation pointer. At different points in time, the concentration of news events varies. A histogram portrays a similar picture, in a plot when the user clicks the History tab.

4.3.2 3D-Plot: Theme dependent – Strength vs. time vs. locations

Figure 12 shows the map of a particular event and its spread over time and locations. An event is typically focused at a particular location, however, as time elapses, it is seen that a community of interest grows that is dispersed in space. This linkage is better presented on a map. The advantage of using the approach discussed in the paper is 1) to explicitly show the time and geospatial evolution of content; 2) to characterize or estimate the spread of a particular event easily. Figure 12 shows that topics that surround the situation in Iraq have global impacts, particularly in Washington, DC, Amman, Jordan and Denver, Colorado. GeoTracker allows us to analyze the correlation of such events.

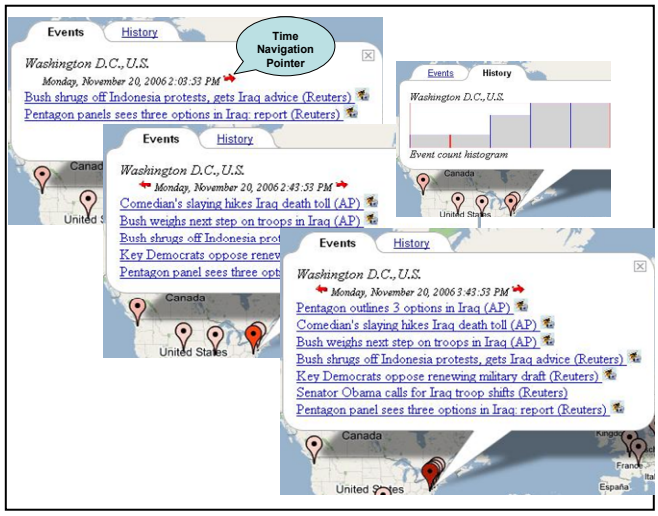

Figure 11 – Time evolution of events fixed at a location

4.3.3 Theme dependent – Strength vs. time

In addition to the spatial representations, a temporal representation can be generated for archival content to give users a sense of how a particular topic of interest is covered by the media over time. Figure 13 shows the results of a query for "World Cup Soccer" displayed as a histogram, and clearly shows the relative media coverage of the topic during the World Cup years, and in between. The histogram bars are hyperlinked to query the database with the additional temporal restriction specifying the selected year.

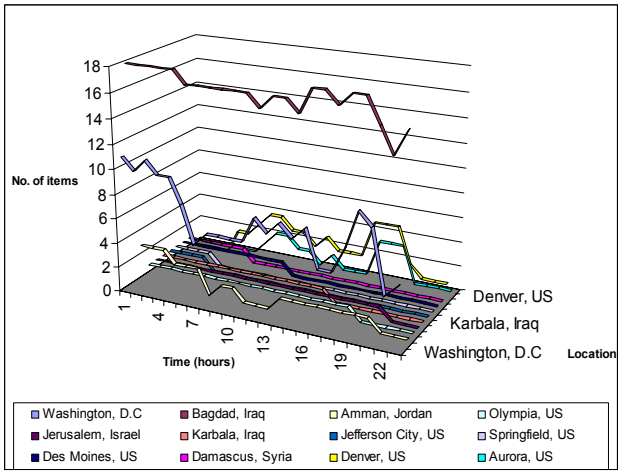

Figure 12 - A particular event (Iraq war) tracked over time and location

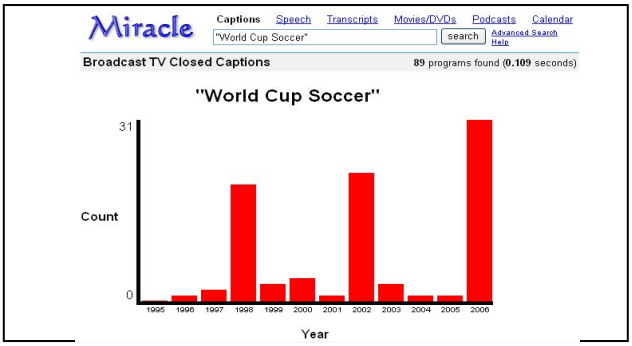

Figure 13 - Strength vs. time for a world sporting event (World Cup soccer)

4.4 GeoTracker performance evaluation

As briefly described in section 2.1, extracting and geo-coding RSS feeds in GeoTracker involve mainly three steps: a) location extraction; b) location mapping; c) geographical coordinate resolution. While steps b) and c) are constrained by the coverage of the location database resources, step a) depends on the effectiveness of the location extraction algorithm implemented in GeoTracker. To evaluate the performance of the overall system, we mainly considered two different approaches for a). First, we implemented an efficient rule-based tagger (RBT) which parses the sentences from the RSS text and extracts a set of locations based on regular expressions and a heuristic algorithm designed around best practices. This provided easy integration and light weight processing suitable for large scale applications. Second, we tested a more sophisticated named entity recognizer (NER) based upon a regularized maximum entropy classifier with Viterbi decoding. This approach is similar to the one described in [31] and has been trained with the CoNLL-2003 [24] annotations and several other syntactic, lexical and word features (e.g., Part of Speech tags, Noun Phrase chunking, capitalization formats, gazetteer words presence, etc.) extracted from the Reuters corpus. A more detailed description of the algorithm has been submitted for publication.

We used three test sets. The test sets A and B consist of the two publicly available data sets used in the CoNLL-2003 language-independent named entity recognition task. Both files are extracted from the Reuters domain, manually annotated and made available to the research community. The test set C has been derived from Yahoo news feeds, covering different subjects (news, politics, sports, etc.) and including 139 items, each one composed of the *title* and the *description* RSS item elements. It was manually annotated by the authors.

4.4.1 GeoTracker geocoding evaluation

In this section, we quantify the accuracy of the GeoTracker geocoding. Figure 14 illustrates the framework for comparing the performance of GeoTracker. The *Upper Bound* performance is provided by passing the annotated (location entities only) files through the GeoTracker Mapper functionality. This provides the truth with respect to precision. The NER classifier has been first trained and subsequently used to perform named entity recognition on the sample files A, B and C. The recognized locations are then passed through the Mapper to obtain a comparison metric. Finally, GeoTracker performs name recognition using the RBT module on files A, B, and C followed again by the Mapper. Note that GeoTracker does not perform any learning and hence does not need any training.

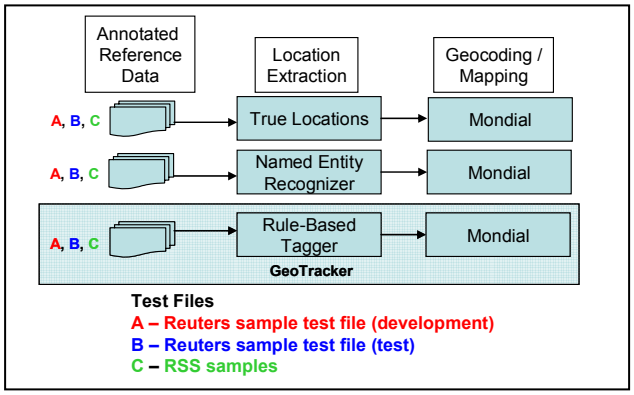

Figure 14 - Experiments framework

The annotated reference data sets are used to establish the upper bound performances. The annotated locations are extracted and mapped through the Mondial database. Locations such as names of continents, non political regions, or landmarks (i.e., Africa, Balkans, Wall Street), are typically filtered out. This gives the maximum number of *mappable* locations (see *Upper Bound* column in Table 1 and Table 2) based on the content of the Mondial data.

The test sets are then passed to the NER module and the recognizer locations are mapped again. We proceed similarly for the RBT module. For each set we computed precision, recall and F-measure, defined respectively as $P=tp/(tp+fp)$, $R=tp/(tp+fn)$ and $F=2PR/(P+R)$, where *tp* are the true positive locations, *fp* the false positive and *fn* the false negative. Results are summarized in Table 1 and Table 2

Table 1. Named Entity Recognizer Performance

Test Set / # locations	Upper Bound	Coverage	NER			
	Mapped / # locations	Mapped/ Mappable	Mapped/ *tp* locations	P (%)	R (%)	F
A / 1829	81.36% 1488/1829	98.34% 1464/1488	84.33% 1464/1736	92.24	93.85	93.04
B / 1662	82.31% 1368/1662	94.30% 1290/1368	86.81% 1290/1486	87.94	87.94	88.33
C / 130	71.54% 93/130	96.78% 90/93	73.17% 90/123	84.62	98.08	88.64

The NER system achieves high values of precision and recall on both test sets from the Reuters domain. This is expected since it has been trained with data from the same domain. It also does quite well with the RSS feed domain, retrieving 123 of the 130 locations (R=98.08%) present in the data, although the precision reduces to 84.62%. When we consider the mappable locations, NER is very close to the upper bound mapping 90 of the 93 mappable locations from the test set C. Overall, the obtained coverage, defined as *Coverage=mapped/mappable* locations, is very close to the upper bound.

Table 2. Rule-Based Tagger Performance

Test Set / # locations	Upper Bound	Coverage	RBT			
	Mapped / # locations	Mapped/ Mappable	Mapped / *tp* locations	P (%)	R (%)	F
A / 1829	81.36% 1488/1829	95.26% 1418/1488	80.66% 1418/ 1758	25.29	90.64	39.55
B / 1662	82.31% 1368/1662	97.95% 1340/1368	82.26% 1340/ 1629	23.98	92.93	38.12
C / 130	71.54% 93/130	98.92% 92/93	71.32% 92/ 129	12.64	96.15	22.34

The RBT system shows a much lower precision (from 12% to 25%), but captures adequately most of the locations (recall ranging from 90% to 96%). The filtering effect of the Mondial

database improves drastically the performances extending the coverage to almost 99% in the case of RSS test set.

4.4.2 Comparison and discussion

Our observations in the previous section show that a simple rule-based tagger like GeoTracker is adequately accurate and sufficient for the purposes of geolocating RSS feeds. The precision ranges from 12% to 25% but the recall percentage is very high. NER classifiers on the other hand often use sophisticated models that require extensive training to achieve high precision and high recall percentages. GeoTracker does not utilize any training. It is important to point out the following characteristics:

- Both approaches implicitly acknowledge that location information is derived from names
- In the NER case, names are selected through a trained model; in our case, through a fairly simple grammar
- For NER based classifiers, names that carry location information are determined by understanding context, hence, requiring sophisticated techniques; GeoTracker is context independent and extracts locations by matching against a database of known locations (i.e., Mondial). We guarantee geocoding for all locations.
- GeoTracker resolves locations to an address format that can be easily geocoded; NER classifiers would need an additional step to realize this.

It is also worth mentioning that GeoTracker can improve the precision by using additional dictionaries that carry alternative names (e.g., N.J. or NJ for New Jersey)

5. Related Work

Geomapping RSS feeds: There has been renewed interests in geomapping content especially photos (e.g., Flickr and other third party service providers like Smugmug [7] and MapBureau [8]). There are many benefits to geotagging content. Geo-located information focuses the interest of the audience to answering questions like "Where did that happen?", "Was it near a special landmark?" or "Was that event near where my colleagues/friends live?" Geotagging any kind of data is now starting to be provided as a service (see MapBureau [8] or GeoUrl [9]). In particular, the RDFIG Geo vocabulary from W3C [27] is the common basis. It supplies official global names for the latitude, longitude, and altitude properties. Providers like Google and Yahoo provide APIs to help geotag content efficiently. In this paper, we have investigated how RSS feeds should be geolocated in an automatic manner. GeoRSS [17] is an RSS namespace extension for encoding location in RSS. CMU's Informedia project has also investigated integrating spatial mapping with video indexing systems [15]. None of these focus on geomapping RSS feeds.

Time Mapping RSS feeds: The other aspect visited in this paper is the chronological or temporal presentation of RSS feeds. This allows us to answer questions like "What are the most prominent events over the past 10 days?" or "Show me the shipwrecks over the past 10 years along the mouth of Columbia River". This concept has also been studied by mapping organizations (see MapBureau [8]. Visualizing tags over time was also presented in [1] whereby the authors devised novel algorithms and data structures to characterize the most interesting tags associated with a sliding interval of time. Google Trends [23] allows users to view the popularity of search items over a period of time. For example, it shows that the keyword YouTube enjoys increasing volume and popularity in the past 12 months. In this paper, we study the timely evolution of events captured in RSS feeds and their changing geographical distribution. In addition, we present the novelty of the algorithms (aggregation, search and personalization) and data structures to efficiently generate this visualization in real-time. Model construction for mining spatiotemporal theme patterns from weblog data was also investigated in [6]. The authors use a probabilistic approach to model the subtopic theme and spatiotemporal theme patterns simultaneously.

6. FUTURE WORK

GeoTracker, while effective in showing the spread and concentration of events in a particular region, is not effective in showing clusters on the same topic that might be distributed in different regions in the world. This latter capability is being explored as part of our future work where topics are clustered using various multi-dimensional algorithms and visualized on an image map. The combination of topic clustering on an image-map and the GeoTracker capabilities may offer the best of both worlds.

GeoWiki – Much work remains to be done on automatically geolocating any named entity types (e.g. person, organizations etc.). In this context, our future work will involve constructing Wiki pages that provide geocoded information for any named entities. GeoTracker can utilize these Wiki pages as additional information to geolocate entities. The hope is to construct a social network of geocoded pages that are continually verified and updated by the community.

User intentions in web applications have been studied from different perspectives [30]. A direct extension of our work is to analyze user intentions at using GeoTracker and provide additional personalized content to further satisfy user's browsing need.

7. CONCLUSIONS

Geolocating content in general is a challenging task. In this paper, we investigated the spatio-temporal mapping of RSS items. In particular, we developed a system that is capable of integrating RSS feeds from publishers and users alike and displaying them on a map. We also demonstrated how a user can easily navigate the time and space evolution of breaking events. The navigation can be performed on a corpus of collected data to revisit/summarize some of the prominent events that have occurred throughout the collection of events, or through a tracker that monitors live RSS feeds. Publishing or retrieving events from mobile devices and the timely dissemination of information to them are also investigated. A middleware platform was introduced to facilitate system integration of the various required components. Finally, we compared the precision and recall metrics for GeoTracker that employs a simple rule-based tokenizer with sophisticated NER based classifiers. Our observations conclude that GeoTracker performs accurately with RSS feeds and do not motivate the need for NER based classifiers. It is our belief that the kind of spatio-temporal navigation of RSS feeds presented in this paper is an important paradigm shift in browsing updates on any topic of interest. It provides users new perspectives on the evolution of events and topics that cannot be easily made available in previous browsers.

8. REFERENCES

[1] Hammersley B., "Content syndication with RSS", O'Reilly, April 2003.

[2] Dubinko M., et al., "Visualizing Tags over Time", In WWW '06: Proceedings of the 15th international conference on World Wide Web (Edinburgh, Scotland). ACM Press. New York, NY, USA, 193--202.

[3] Web 2.0 - http://en.wikipedia.org/wiki/web_2.0

[4] Chen, Y-F., et al. "iMobile EE - An Enterprise Mobile Service Platform," Wireless Networks 9(4): 283-297. 2003.

[5] Dennis, B. and Jarrett, A. "NusEye: Visualizing Network Structure to Support Navigation of Aggregated Content," Computer Science Department, Northwestern University. Hawaii International Conference on System Sciences (HICSS-38), Persistent Conversation Minitrack: 2005, to appear , .

[6] Mei, Q.,et al,"A probabilistic approach to spatiotemporal theme pattern mining on weblogs", In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM Press, New York, NY, 533-542.

[7] SmugMug Inc., "SmugMaps: combining the power of Google Maps with 65,000000+SmugMug photos",2006,http://maps.smugmug.com/

[8] http://www.mapbureau.com/geotagnow.html

[9] GeoURL (2.0). 2006. http://geourl.org/

[10] Digg Inc. 2006. http://digg.com

[11] Flikmap. A Semsym project. 2006. http://flickmap.semsym.com/

[12] Liu Z., et al. "Multimedia Content Acquisition and Processing in the MIRACLE System," in IEEE CCNC, Jan. 7, 2006.

[13] Gibbon, D.et al. "The MIRACLE Video Search Engine," in IEEE CCNC, Jan. 7, 2006.

[14] Clinton, D. "OpenSearch Specifications, 1.1, Draft 3," A9.com, http://www.opensearch.org/Specifications/OpenSearch/1.1

[15] Christel, M. et al., "Interactive Maps for a Digital Video Library", IEEE Multimedia, 7(1), pp. 60-67,, 2000

[16] Pan, J.Y., Faloutsos, C, GeoPlot: Spatial Data Mining on Video Libraries, in Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM'02), Mclean, Virginia, Nov. 4-9, 2002

[17] GeoRSS http://www.georss.org

[18] Yahoo. "MediaRSS – Specification Version 1.1.1", Oct. 2005. http://en.wikipedia.org/wiki/Media_RSS

[19] May, W. "Information Extraction and Integration with Florid: The Mondial Case Study," Dec. 1999. The Mondial Database: http://www.dbis.informatik.uni-goettingen.de/Mondial/

[20] Wikipedia, http://en.wikipedia.org/wiki/Wiki.

[21] Flickr. "Explore everyone's geotagged photos on a Map", http://www.flickr.com/map.

[22] Nottingham, S. and Sayre, R. "The Atom Syndication Format", IETF RFC 4287, Dec. 2005. http://tools.ietf.org/html/rfc4287

[23] Google Trends, http://www.google.com/trends

[24] Reference labeled data for Named Entity recognition, http://www.cnts.ua.ac.be/conll2003/ner/

[25] Reuters corpus - http://about.reuters.com/researchandstandards/corpus/

[26] W3C Mobile Web Initiative Device Description Working Group, http://www.w3.org/2005/MWI/DDWG/.

[27] W3C RDFIG Geo vocabulary, http://www.w3.org/2003/01/geo/.

[28] Google Map API - http://www.google.com/apis/maps/

[29] Wei B. et al. , "MediaAlert—A Broadcast Video Monitoring and Alerting System for Mobile Users," The 3rd International Conference on Mobile Systems, Applications, and Services (MobiSys), Seattle, June 2005.

[30] Hiramoto R. and Sumiya K., "Web Information Retrieval Based on User Operation on Digital Maps",14th International Symposium on Advances in Geographic Information Systems, Nov. 2006

[31] Dudik M.et al.,"Performance guarantees for regularized maximum entropy density estimation", Proceedings of the 17th Annual Conference on Learning Theory, 2004, 472-486