Geospatially Visualizing Users' Preferences from APML Profiles

Timothy Schultz, Robert Reber, Heather White - Drexel University Masters of Information Systems

Abstract - *Data mining and analysis of user information is not a new technology. Many techniques have been developed in order to capture the habits and interests of the user. Speculation on users habits have also been extracted from such sources as click streams, databases and other user activities. Each of these methods leverages the implicit knowledge about a user's historical and habitual actions. Suppose information could be polled from the user in various abstracted ways, such as a questionnaire, which can pinpoint explicit concepts about a user's preferences. What we plan to introduce is a new concept based on visualization. Geospatial APML mapping is a new unifying framework geared at geospatially visualizing attention information and its changes over time. This paper looks to demonstrate the feasibility of presenting this information via a Google Earth geographical representation of a user's attention profile. The paper also shows other potential uses for this type of data mining when used with the APML open standard. Our hope is that by conducting research and analysis of the outlined concept, an even greater potential for benefits may be realized.*

Index Terms - geospatial mapping, information visualization, APML, Google Earth, Attention Profile Markup Language

1. Introduction

Analyzing users' behavior on the Web is a very important and challenging research topic. If users' interests can be automatically detected it can be used for information recommendation and marketing which are useful for both users and Web site developers [1]. The importance of this data is prevalent and can be seen in the many techniques used trying to capture it. A characteristic of most of the existing approaches in obtaining data is that these methods build profiles only on data collected by a single web site about the users' access, called site-centric data [3]. In a study done by Padmanabhan on site-centric and user-centric data collection, Padmanabhan found user-centric data outperformed models built on site-centric data by 2-1. The results provide evidence that conclude site-centric data alone prove to incur potentially erroneous conclusions [3]. Perhaps this study is telling us that understanding users' habits does not present a clear picture of understanding the user.

What is needed is a way to move from the site-centric to a user-centric technology. Until now that method has not been widely available. With the advent of APML, the Attention Profiling Mark-up Language, this methodology can easily be adapted and used interchangeably across sites and between vendors.

What is attention data? Attention data is comprised of the websites you visit, the things you might write about in your blog, the music you listen to through such services as Last.fm, the websites you bookmark using social tools like del.icio.us, the photos and videos you share with flickr and YouTube [4]. Another way of looking at an attention profile is:

"*...consolidated, structured descriptions of people's interests and dislikes. The information about your interests and how much each means to you (ranking) is stored in a way so that computers and web-based services can easily read it, interpret it, process it and pass it on should you request and permit them to do so.*" [5]

APML provides a standardized way to collect and rank your attention data [4] through sets of domain-specific concept and value pairs. Each attention concept has a corresponding double value between zero (low) and one (high) which indicates the user's affinity to that concept. A standardized protocol such as APML can take user data mining from a site-centric model to a user-centric model, as any allowed vendor deploying APML services can write implicit concepts to a user's profile.

2. Conceptualization

Collecting the data is only the first part. Getting the data into the hands of someone that can use it is another issue. Information Visualization can quickly highlight an area that may require closer interrogation. Information visualization also holds the promise of alleviating information overload on the Web by summarizing a large amount of data into a two-dimensional display format such as a map [2]. Visualization can quickly assist the market in responding to consumer desires while making the best use of distribution and storage facilities. Empirical studies have proven that visualization in a geospatial context have been beneficial in relaying information quickly and easily. FEMARepViz is one program that uses Google Earth to enhance visualization. Processed reports are stored in a repository and can be retrieved by a web interface. The output visualization of processed report data is a KML document that provides dynamic updates, interactive visualization, and a query CDA link. Users can click on the query CDA link to retrieve local news stories through GN that are relevant to the incident that was extracted by FEMARepViz [4]. This use of visual technology has the backing of the National Science Foundation along with The National Visualization and Analytics Center. Backings by such prestigious organizations validate the concept and provide an excellent foundation for extended research in this domain.

Our concept uses the example of tracking the various user preferences of wine. This of course is one of a myriad of possible scenarios. We first query APML data to retrieve references to wine. The data collected will be placed in a data base where it will be combined with latitude and longitudinal values to then be plotted geospatially and weighted by importance as an overlay to Google Earth.

3. Adaptation of the Concept

In this visualization framework Google Earth was chosen as the base application to overlay the captured attention data. Google Earth is a familiar tool used for viewing geospatial information. It is a searchable 3-dimensional combination of maps and imagery that depicts the planet and surrounding space. The group that initially built the application which has since been acquired by Google developed an XML standard for data to be overlaid on the mapping application called Keyhole Markup Language KML. KML can be used alone or can be combined with other media into archive files called KMZ. This framework uses the KML format to map the data compiled from the users APML profiles.

The purpose of this visualization is to show which geographical regions a users wine preferences are linked to by overlaying that data onto the Google Earth application.

For this dataset we have three zoom levels of mapping. They are by country (Green), region (Purple), and sub-region (Yellow). Each level is rendered as a transparent 3-dimensional box that grows outward into space to show its "importance" to the user. In this case "importance" is based on the quantity of wine owned by the user from each location. The levels are color coded so they can be viewed together or hidden to view any level on its own. This functionality is available in the Places pane in Google Earth.

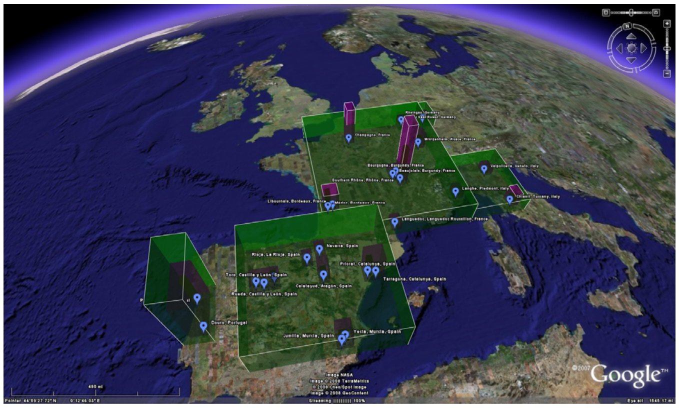

Rheingau, Germany
Saar Ruwer, Germany
Champagne, France
Wintzenheim, Alsace, France
Bourgogne, Burgundy, France
Beaujolais, Burgundy, France
Southern Rhône, Rhône, France
Valpolicella, Veneto, Italy
Langhe, Piedmont, Italy
Chianti, Tuscany, Italy
Libournais, Bordeaux, France
Médoc, Bordeaux, France
Languedoc, Languedoc Roussillon, France
Navarra, Spain
Rioja, La Rioja, Spain
Priorat, Catalunya, Spain
Toro, Castilla y León, Spain
Tarragona, Catalunya, Spain
Calatayud, Aragón, Spain
Rueda, Castilla y León, Spain
Douro, Portugal
Jumilla, Murcia, Spain
Yecla, Murcia, Spain
490 mi
©2007 Google
Figure 1

Figure 1 depicts the view with all levels on. It is easy to discern the importance of country, region, and sub-region as importance has been intuitively represented as a 3-dimensional graph extending away from the surface of the earth creating what we like to call EarthBurst. By allowing this utilization, the user may select to research the data based on the country or region view, as seen in figure 2.

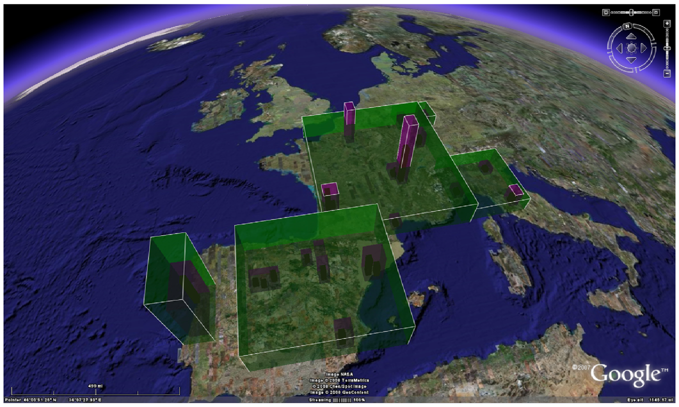


Figure 2

As the level may be controlled with Google Earth, the issue with focus over context can be manipulated easily. Turn off one level to find the most popular wine and its location. As we can see in figure 3, by turning off the other levels we can get a clear understanding of the most important regions according to our search results.

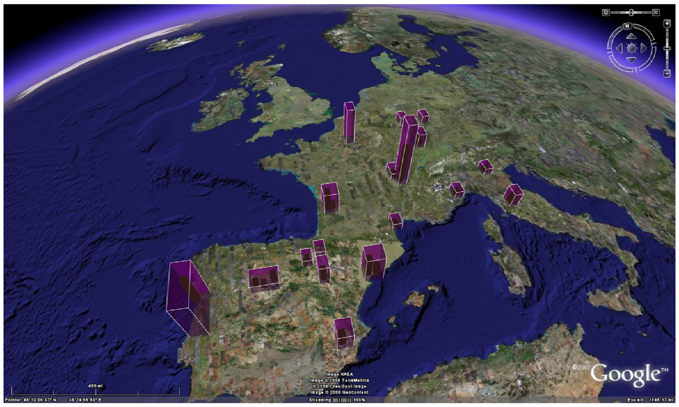


Figure 3

Turn off all levels except sub-levels to find the most popular sub-region, as in figure 4.

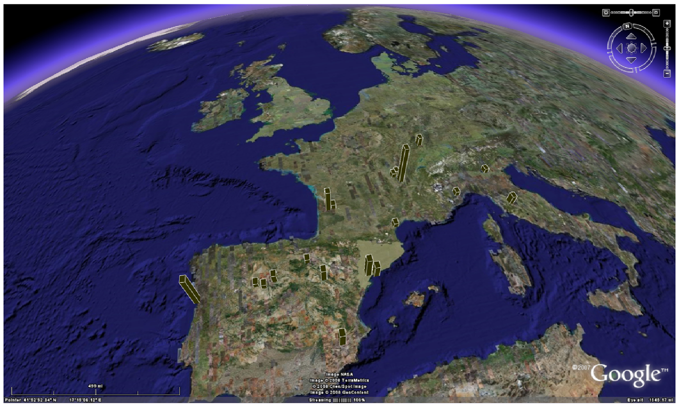
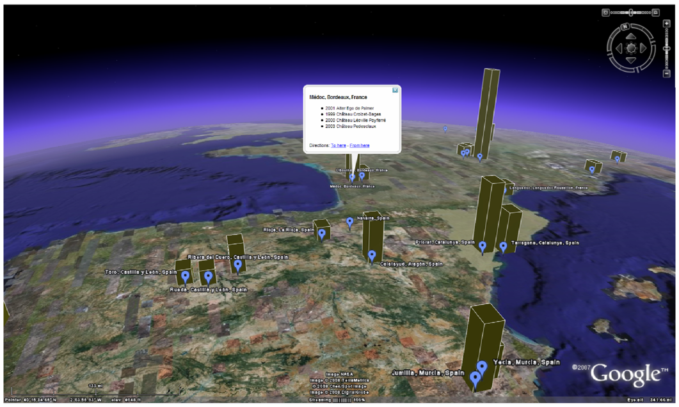


Figure 4

The investigator may find the most popular wine is from Portugal, but the most popular region may be Spain, while at a different level shows the concentration of popular wines resides in the sub-region of France. Again, turn on only sub-region and the most popular wine region from our query in from Chile, as seen in figure 6. This information is easily found by turning on or off different layers.


Médoc, Bordeaux, France

- 2001 Alter Ego de Palmer
- 1999 Château Croizet-Bages
- 2000 Château Léoville Poyferré
- 2003 Château Pedesclaux

Directions: To here - From here

Libourne, Bordeaux, France
Médoc, Bordeaux, France
Languedoc, Languedoc Roussillon, France
Navarra, Spain
Rioja, La Rioja, Spain
Priorat, Catalunya, Spain
Tarragona, Catalunya, Spain
Ribera del Duero, Castilla y León, Spain
Calatayud, Aragón, Spain
Toro, Castilla y León, Spain
Rueda, Castilla y León, Spain
Yecla, Murcia, Spain
Jumilla, Murcia, Spain
©2007 Google™

Figure 5

Upon clicking a specific place marker, concepts derived from the user's attention profile are linked to that specific location in the form of an unordered list as depicted in figure 5.

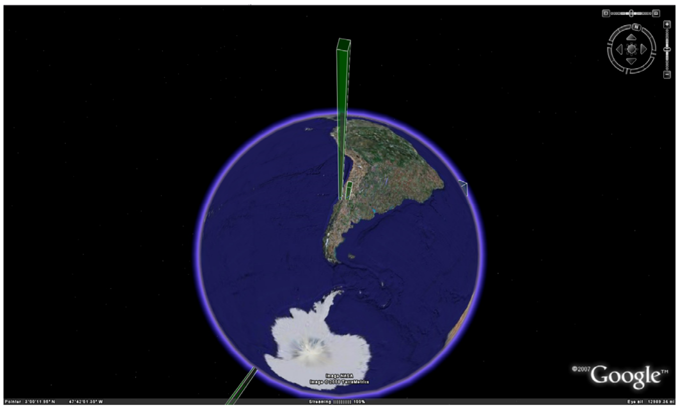


Figure 6

The ability to trend this information over time is also a very powerful tool. [7] To see the importance of a region drop and a sub-region to become very important may point to a climate change for that year, or a new vintage from the grower, or even a new technique in pruning or irrigation. To be able to see all three regional levels at once enables the user to both understand the overall picture and the finer granular details simultaneously. This method of encapsulating children regions within parent regions also establishes a hierarchical structure that is easily visible upon viewing the visualization.

4. Use Case – Wine Mining

The overall goal for this case is to analyze patterns and trends in wine preferences by the study of APML profiles based on user purchases and reviewing activities. Wine enthusiasts are constantly discussing where the next "big" wine producing regions will be, both overall and for specific types of wines. For example, some people may fervently argue that Pennsylvania Cabernet Francs will emerge as an incredibly popular wine within the next ten years. This was a primary reason for using wine preferences as an example use case; such topics are generally open for much debate, and can be visually validated with our proposed framework.

Although not all the information will be used in this case study, it should be known that the ability to map a specific location to a general point on a map, map a specific wine to its location of origin and map a users ranking of specific wines are part of the functionality. There are also extended capabilities such as price, ratings and production levels that could be used if desired.

Interesting information can be obtained from these queries once visualized. A quick query may tell you what the most popular wine is at that given time. You can immediately see the geographic locations of the wine regions plotted on the map and their importance.

This information in the future can be viewed over time to see if there are shifts in either location or importance. Incremental snapshots of a user's APML profile can be visually rendered and merged to create an animation of preference changes over time. By including temporal mapping the point of origin for a person's preference, in this case wine, you can begin to see trends on region preferences emerge.

By combining queries new and exciting trends may be discovered. Perhaps a query on APML profiles of southern US states turns up a trend towards Australian wines by that section of the country? Again, visualization will present this information very quickly. Or perhaps a query on Pennsylvania Cabernet Franc shows that over time it is trending upward.

5. Future Applications

The framework discussed here is easily adapted to many datasets. There are numerous preferences that can be mapped to specific geographic locations. For example travel and leisure activities, food and beverages tastes, news outlets, pretty much anything that can be mapped to a location.

There is also the ability to use this framework as a geodemographic system that can take data from multiple sources to visualize where users with similar preferences are located in proximity to each other. This helps determine markets where like-minded people may live.[6] Using the wine example, a distributor may notice a trend that people in a certain geographic area share a preference for wine produced from a specific regions. They can then adjust their distribution accordingly to account for those preferences.

6. Technical Overview

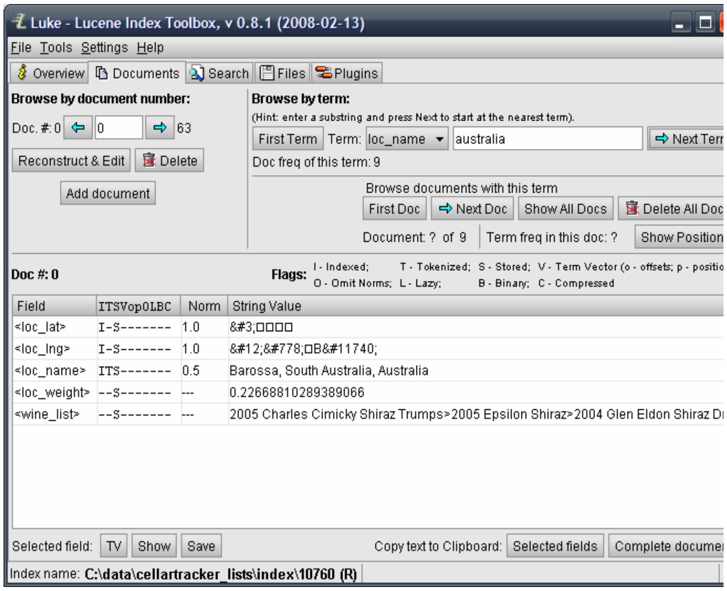
To develop the proposed solution, numerous open source Java libraries were utilized. To better understand the technologies behind the specific use case outlined in this paper, it is appropriate to analyze the steps taken to generate the final visualizations. It will be noted that we plan on releasing this framework under an appropriate open source license in the near future.

1. Data Acquisition

Since APML is a relatively new specification, we were unable to locate any wine vendor websites that adopted APML exports with their services. We could, however, locate a handful of sites where users publicly displayed a maintained list of their currently purchased and reviewed wines. We downloaded a sample of 100 publicly available wine profiles in HTML format, applied a regular expression to the contents, and filtered out any HTML markup within the text. From here, we were able to parse out all the relevant information contained within each user's list (wine name, location, quantity, etc.) Any information retrieved about a unique wine was persisted into a *Postgres* database for later retrieval. *APMLStream* was used to write the raw data to an APML object that was then serialized to the compliant XML-based APML 0.6 specification and written to the hard drive. Two profiles were created for each user APML file: one to store information about "location concepts", and the other to store information about specific "wine bottle concepts". Location concept values were initially calculated by a simple averaging algorithm of wine bottles per location. Wine bottle concept values were also calculated with a simple averaging algorithm based on the quantity of bottles purchased compared to the overall number of wines in the list. Please note that upon greater adoption of APML exporting features between web services, this step will likely not be needed; it was mainly a technique we used to create a test dataset of implicit user preferences. In the future, we foresee this process simplified to acquiring public APML profiles directly from a vendor.

2. Data Persistence

With the test APML files created and stored on the file system, we could then begin the process of linking location and wine bottle concepts to actual locations on a map. Again using *APMLStream* to deserialize each user attention profile to memory, we first iterated through all location concepts stored within the location profile. Numerous pieces of data must be merged with the location concept to appropriately render it within the visualization, such as latitude and longitude coordinates, its APML value, and the list of its associated wines. Because of this, a secondary medium would have to be utilized to tie all of this relevant information together. We chose to use *Apache Lucene* to index these concepts and link together the required information for rendering the visualization. This also enabled the possibility of leveraging Lucene's extensive querying features for filtering out concepts to render customized and specific visualizations. *GeoGoogle* was utilized to directly query Google services for latitude and longitude coordinates of a given location. These coordinates were indexed along with the location concept using *LocalLucene* – an extension to Lucene which enables geospatial queries to be applied to an index, such as retrieving results only within a certain distance to a given coordinate. Again, this is another feature that can be utilized to filter information for more specific visualizations. Lastly, the list of specific wine concepts were matched with their originating location concepts and added to the index. Indexes can be regenerated on an incremental basis to reflect changes in the user's attention profile. The figure below shows the composition of a sample index viewed with Luke, the Lucene index toolbox.



Browse by document number:

Doc. #: 0 ⇦ 0 ⇨ 63

Browse by term:

(Hint: enter a substring and press Next to start at the nearest term).

Term: loc_name ▾ australia

Doc freq of this term: 9

Browse documents with this term

Document: ? of 9 | Term freq in this doc: ?

Doc #: 0

Flags: I - Indexed; T - Tokenized; S - Stored; V - Term Vector (o - offsets; p - positio
O - Omit Norms; L - Lazy; B - Binary; C - Compressed

Field	ITSVopOLBC	Norm	String Value
<loc_lat>	I-S-------	1.0	&#3;□□□□
<loc_lng>	I-S-------	1.0	&#12;&#778;□B&#11740;
<loc_name>	ITS-------	0.5	Barossa, South Australia, Australia
<loc_weight>	--S-------	---	0.22668810289389066
<wine_list>	--S-------	---	2005 Charles Cimicky Shiraz Trumps>2005 Epsilon Shiraz>2004 Glen Eldon Shiraz Di

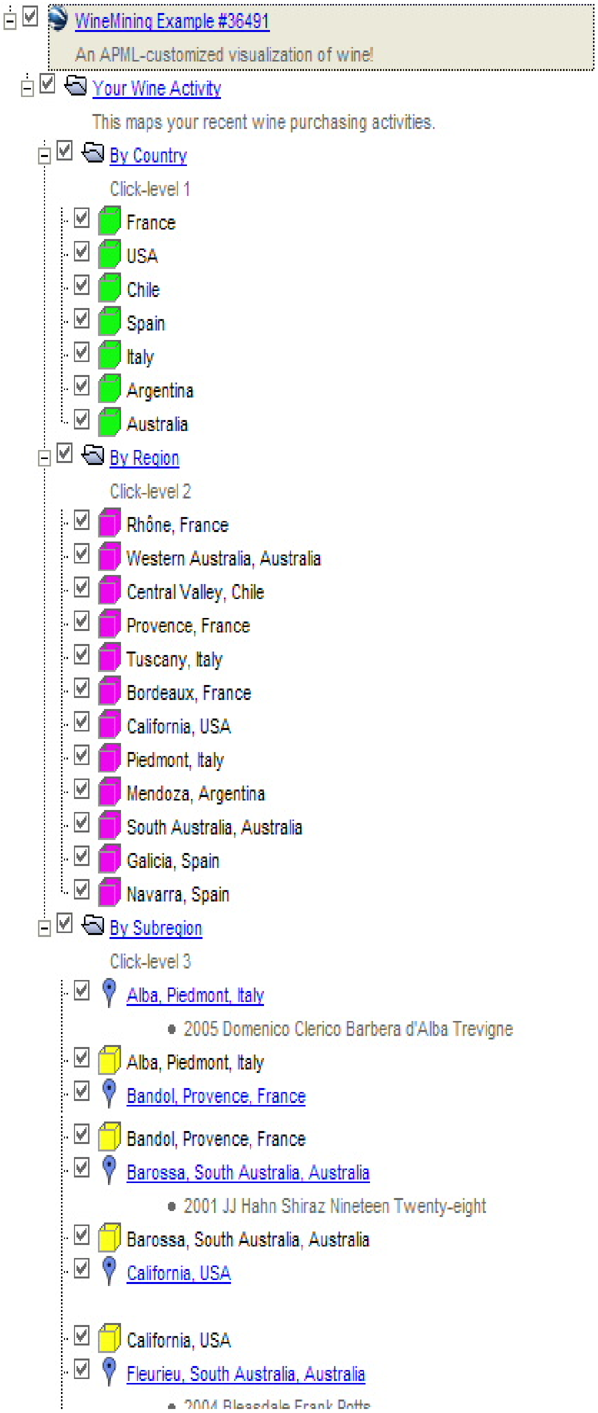
Selected field: Copy text to Clipboard:

Index name: **C:\data\cellartracker_lists\index\10760 (R)**

Analyzing a sample user index
Figure 7

3. Data Conversion

Having created the user indexes with the appropriately linked information, attention data can now be extracted for rendering in Google Earth. Using Lucene's extensive querying features, information can be filtered for customized visualizations. In the examples provided in this paper however, an overall view of the collected data is being considered. Because of this, an all inclusive query is being applied. Iterating through these indexed documents, information is pulled from each field (loc_lat, loc_lng, loc_name, loc_weight, wine_list), and appropriately applied to various KML objects using *GEKMLLib*. Sub-region placemarkers (level 3) are first generated according to the coordinates found in the user index. Then an algorithm is applied to cluster related sub-regions together in order to form the boundaries of the region (level 2) boxes. The same algorithm is then applied to cluster related regions together for the creation of the country (level 1) boxes. This produces the "encapsulated" feature demonstrated in the visualizations in which a parent level encompasses all children levels. Box height for each level is calculated based on its ranking within the user's attention profile. The KML object is then serialized and stored on the file system for viewing in Google Earth. The layer-upon-layer effect can be manipulated in Google Earth's side panel, as illustrated.



- ☑ WineMining Example #36491
  An APML-customized visualization of wine!
  - ☑ Your Wine Activity
    This maps your recent wine purchasing activities.
    - ☑ By Country
      Click-level 1
      - ☑ France
      - ☑ USA
      - ☑ Chile
      - ☑ Spain
      - ☑ Italy
      - ☑ Argentina
      - ☑ Australia
    - ☑ By Region
      Click-level 2
      - ☑ Rhône, France
      - ☑ Western Australia, Australia
      - ☑ Central Valley, Chile
      - ☑ Provence, France
      - ☑ Tuscany, Italy
      - ☑ Bordeaux, France
      - ☑ California, USA
      - ☑ Piedmont, Italy
      - ☑ Mendoza, Argentina
      - ☑ South Australia, Australia
      - ☑ Galicia, Spain
      - ☑ Navarra, Spain
    - ☑ By Subregion
      Click-level 3
      - ☑ Alba, Piedmont, Italy
        - 2005 Domenico Clerico Barbera d'Alba Trevigne
      - ☑ Alba, Piedmont, Italy
      - ☑ Bandol, Provence, France
      - ☑ Bandol, Provence, France
      - ☑ Barossa, South Australia, Australia
        - 2001 JJ Hahn Shiraz Nineteen Twenty-eight
      - ☑ Barossa, South Australia, Australia
      - ☑ California, USA
      - ☑ California, USA
      - ☑ Fleurieu, South Australia, Australia
        - 2004 Bleasdale Frank Potts

Side-panel demonstrating the three "click-levels"
Figure 8

7. Issue to be Considered

At the writing of this paper APML is still a new concept. No empirical studies have been performed that can be referenced to show the degree of success of APML. Success will be gauged by future vendor adoption of the technology. APML is an open source initiative that is being looked at by corporations such as Microsoft and

Google. For this paper a limited dataset was created of 100 APML profiles as to show the validity of this research. The Use-Case and visual representations clearly show and warrant the future studies of this framework.

8. Conclusion

Since this is our initial research of a relatively new technology, it would be appropriate to discuss some general issues and lessons gained from this paper. Firstly, since APML is a new open standard, there are currently no empirical studies that could be found geared specifically towards the feasibility of visualizing APML profiles. Because of this, we've established our own framework of acquiring, storing, rendering, and visualizing attention data. We will continue to develop these newly formulated ideas and incorporate new APML specification features as they become available. Also, it is realized that the success of such a framework greatly relies on the adoption of APML within domain-specific web services. Currently, the visualizations presented were derived from implicit concepts only. It will be interesting to compare these visualizations to a more accurate explicit concept dataset once they become available. We must also consider ways to accurately aggregate APML information between different vendors. Since numerous sites can write to a single APML profile, each vendor may have its own way of calculating concept values. These differences will have to be accounted for in the future when rendering our proposed visualizations. Lastly, we will have to continue the validation of our results. As expected, the initial data extracted from the users' public wine lists had to be groomed – especially location names so they were recognized by Google's geo service. It would be wise to continue this validation with willing domain experts.

References:

[1] [1] Tsuyoshi Murata, Kota Saito, Extracting Users' Interests from Web Log Data Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence Pages 343-346 Year of Publication: 2006 ISBN:0-7695-2747-7 Publisher IEEE Computer Society Washington, DC, USA

[2] Wingyan Chung, Ada Leung Supporting Web Searching of Business Intelligence with Information Visualization Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence Pages 807-811 Year of Publication: 2007 ISBN:0-7695-3026-5 Publisher IEEE Computer Society Washington, DC, USA

[3] Balaji Padmanabhan, Zhiqiang Zheng, Steven O. Kimbrough, Personalization from incomplete data: what you don't know can hurt Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining San Francisco, California Pages: 154 - 163 Year of Publication: 2001 ISBN:1-58113-391-X Sponsors SIGMOD: ACM Special Interest Group on Management of Data AAAI : American Association for Artificial Intelligence Publisher ACM New York, NY, USA

[4] Brian M. Tomaszewski, Chi-Chun Pan, Prasenjit Mitra M. MacEachren Facilitating situation assessment through gir with multi-scale open source web documents Proceedings of the 4th ACM workshop on Geographical information retrieval Lisbon, Portugal SESSION: Mining geographic information and GIR applications Pages 95-96 Year of Publication: 2007 ISBN:978-1-59593-828-2 Association for Computing Machinery Publisher ACM New York, NY, USA

[5] APML attention profiling mark-up language. Retrieved March 10, 2008, from APML Web site: http://www.apml.org/

[6] Goss, J. 1995. ""We Know Who You Are and We Know Where You Live": The Instrumental Rationality of Geodemographic Systems." Economic Geography 71(2): 171-198.

[7] Weaver, C., Fyfe, D., Robinson, A., Holdsworth, D., Peuquet, D., and A.M. MacEachren. 2007. Visual Analysis of Historic Hotel Visits. Information Visualization 6(1). 89-103.