BetSmart

MIS510 Final Report

Cai Chen, Ximing Yu, Ying Jin
Spring 2010
Table of Contents

1. Introduction to BetSmart ................................................................. 3

2. Site Tour ......................................................................................... 3
   2.1 General soccer fans ................................................................. 3
   2.2 Online Better ........................................................................... 6

3. Business Model ............................................................................... 8
   3.1 Revenue Model ........................................................................ 8
   3.2 Competitive Analysis ............................................................... 8

4. System Architecture ......................................................................... 9
   4.1 Google App Engine ................................................................. 10
      4.1.1 Advantages of GAE .............................................................. 11
      4.1.2 Disadvantages of GAE ......................................................... 11
      4.1.3 Suggestions to using GAE .................................................... 11
   4.2 Data Collection ........................................................................ 12
      4.2.1 Gazzetta dello Sport .............................................................. 12
      4.2.2 BetExplorer ........................................................................ 12
      4.2.3 Football-Italia ..................................................................... 12
   4.3 API Integration ........................................................................... 12

5. Data Mining Component ................................................................. 13
   5.1 Preparation for data mining ....................................................... 13
   5.2 Data mining using Weka ............................................................ 14
   5.3 Evaluation of the Prediction Model .............................................. 14

6. Conclusion .................................................................................... 15

7. Appendix ....................................................................................... 15
   Team Contribution ........................................................................... 15
1. **Introduction to BetSmart**

   Soccer is the most popular sport in the world. The performance of teams and players can emotionalize their fans greatly. As a byproduct, soccer betting is turning out to be a very profitable field. BetSmart focuses on guiding betting in Serie A of Italy, one of the top soccer leagues in the world. The website provides all the information needed for doing betting, including news, club and player information, matches statistics and predictions of coming matches.

2. **Site Tour**

   BetSmart targets at different segments of users: the first segment would be general soccer fans as they can browse latest news, club standings, match results, videos and so on. The second segment would be online betters. It provides prediction of the coming matches based on history data and data mining techniques. Also it displays odds provided by all betting companies worldwide and labels the highest odds for the 3 outcome (home-win, draw and away-win). Additionally, in case experienced betters are interested in statistical data like odds history and history match results, based on which they can make new predictions, the website collects history data for the past 12 years.

2.1 **General soccer fans**

   There are five main categories of information in this website: News, Matches, Standing, Clubs and Statistics (Figure 1).

   ![Navigation of the website](image1)

   In the home page, the website provides latest news and current Standing of Serie A (Figure 2).

   ![Latest news and current Serie A standing in home page](image2)

   By clicking on the news title, users will be directed to the original page containing that piece of news. For the convenience of users, the news pages are all translated into English by Google Translation.
At the bottom of home page (Figure 3) are introduction to the website, information source and site map. The website also provides some small but functions, including a quick search box and an “AddThis” button through which users can share this website on their Facebook or Twitter.

In “News” and “Standing” page, more detailed news and standing information are provided than in the home page.

In the “Matches” page, all matches in this season are listed by weeks. Users can also use the “Go to week” tool to go directly navigate to a particular week. When clicking on the score of one match, users can get the detailed information about that match. The information includes video of the match from Youtube, detailed report about the match, and so on (Figure 4).
On the Clubs page, all the 20 clubs in serie A are listed. By clicking on the logo, users can go to the detail page about that club. Information such as leadership, technical stuff, honors, squad and matches in this season is listed (Figure 5).

Users can also check out the information about one particular player by clicking on the name in the squad list. Photos from Flickr are provided, as well as the player’s information, performance in this season and career paths (Figure 6).
2.2 Online Better

There is a list of matches in the coming week on the home page (Figure 7). By clicking on the match, users can easily go to the page containing the statistics information about that match.

![Figure 7 List of coming matches](image)

The website provides its own prediction about the result (Figure 8).

![Figure 8 Prediction result from the website](image)

Season ranking trend is provided to show the long term performances of the two teams (Figure 9 Season ranking trend). And recent matches result is provided to show the short term performances (Figure 10).
Squad information is also provided for users to check the formation and players information (Figure 11). The website also collects rival history of the two teams since season 98-99 for users (Figure 12). These are all very important for an online better to do the analysis and final decision.
After looking through the information provided by the website, users should be ready to do the betting. Users always want to know the highest odds for each result so that they can gain more. In order to help users save time searching in different websites, BetSmart has collected odds from different betting companies and highlighted the highest odd for each result (Figure 13). Users can simply see the highest odd and do online betting on that company by clicking on the company name.

3. Business Model

3.1 Revenue Model
Betsmart is a place where you can have fun as well as make profit. Therefore, it is expected to be attractive to general users and profit-seekers. Here are several sources that we can make revenue from:

- Betting companies: As users who want to bet on this website will be directly linked to the betting companies, this is definitely bringing in new customers to those betting companies. Plus, when it recommends the best odds for betters, the betters are very likely to choose the
recommended betting companies. Considering all these, the betting companies would be happy
to partnership with Betsmart.

- Advertisements: Because this website is focused on soccer game and soccer betting, the users
  are very likely to be soccer fans and online betters. Therefore, it would be more effective to post
  soccer-related ads in this website than in others.
- Although for now the website is providing the prediction function freely, it can later set up
  membership mechanism which means that only members are open to the most critical content
  such as the recommendation and prediction. This way, membership charge could be another
  revenue source.

3.2 Competitive Analysis

Betsmart is a one-stop site for soccer fans. It provides rich information in multiple formats: reports,
statistics, pictures, videos and so on. Our major competitors defined are: Italian Football, La Gazzetta
dello sport and BetExplorer.

Italian Football: It provides rich information on matches and the history match results. But it does not
contain anything related to betting. Besides, it doesn’t provide any videos of the matches. Basically, it is
a good information source, but not versatile enough.

La Gazzetta dello: It has very good information of the clubs and players in Italian Serie A. Nevertheless,
because it is a newspaper site, it focuses solely on providing information and facts.

BetExplorer: As indicated by its name, it is a professional website for betting as it has rich statistical
history data. But it does not include any other general information that might interest the soccer lovers. The biggest weakness of this website is that it does not analyze the history data, let alone making
predictions.

4. System Architecture

![System Architecture Diagram]

Figure 14 System Architecture
Figure 14 displays the architecture of our system. The whole system is built on Google App Engine and we choose Struts 2 as the web framework. We collected our data mainly from 3 soccer-related websites: Gazzetta dello Sport, BetExplorer and Football-Italia. Intelligent spiders are built to minimize the efforts needed to collect all the information. Open web APIs, such as YouTube, flickr, Google Translate and Google Visualization, are used to provide multimedia and multilingual components to users. To realize match predication, we utilized the Logistic Regression provided by Weka. Following is a guideline for our system architecture part.

Section 4.1 introduces what Google App Engine is and the cost-benefit of using Google App Engine. Section 4.2 describes our sources of data and how to build the spiders for each data source. Section 4.3 briefly covers how to integrate each web API into our system. The data mining component will be discussed with detail in Section 4.

### 4.1 Google App Engine

Google App Engine (GAE) lies in the cloud of Google and it lets users run their web application on Google’s infrastructure for free when the application is under a certain quota. The current quota is 500 MB of storage and 5 million page views a month, which satisfies the needs of our proposed project.

Currently GAE supports two runtime environments, Java and Python. We chose to Java version since members of our group are more familiar with Java. The architecture of Java version GAE is shown in Figure 15. At the core of GAE is a web server, complemented by data access facilities and various other services enabled by Google.

**JPA** (Java Persistence API) specifies the standards for mapping object to tables, irrelevant of the implementation of the relational database. The data storage behind GAE is Google Bigtable, which is
more like a multidimensional map than relational database. This makes the implementation of JPA in GAE a little bit different from common ones. The JPA implementation of GAE is based on DataNucleus.

**Task queues** allow requests to be queued and executed in parallel. Since GAE has a 30-second limit on the execution of time of each request, task queues help a lot when sophisticated transactions are needed. With task queues, developers can divide a big task into several smaller ones and put all them into the queue and GAE will manage that queue for users.

**URL fetch** enables Java programs to create URL objects which can be used to exact contents of other URLs. The spiders for collecting data are based on URL fetch and we will discuss how to build the spiders in Section 4.2.

**Memcache** provides an implementation of Java L2 Cache and the benefits of using Memcache is that developers can store the commonly used data in Memcache (memory) and thus reduce the number of times for data retrieval from data store. In the developing process, we find that using Memcache can tremendously boost the responsiveness of our application.

**Google accounts** can also be integrated into GAE so that the application can authorize users based on their Google accounts. For our application, we use Google accounts to specify the system administrator of the site so that private pages can be properly protected.

### 4.1.1 Advantages of GAE

Based on our experience developing with GAE, we found GAE to be of both advantages and disadvantages in different aspects. The advantages in using GAE are:

- **GAE is free of charge** for small or mid-scale application. It saves our time and money to find a scalable and stable server for hosting the application.
- The project structure of GAE is similar to common web project, so the learning curve is not very steep for those with web development experience on web servers like Tomcat.
- Google provides an Eclipse plugin for developing GAE. The plugin contains an implementation of web server (Jetty) and data store which can run on local developer machines. The debugging tool with the plugin is very useful.
- GAE provides a well-configured JPA implementation and the only codes needed for object persistence is the Java annotations on each entity class.
- GAE provides a lot of handy services, such as Task Queues, URL Fetch, Memcache, etc. GAE is shipped with default configurations for most of the services, so minimum configuration is required.

### 4.1.2 Disadvantages of GAE

The disadvantages with GAE include:

- There are many restrictions on the Java runtime of GAE. For example, the program is not allowed to write files to the server, though the program can read from files transferred to the server. Library for parsing XMLs, like jdom, cannot run on GAE, which results in a lot of pains when processing XML retrieved from other web services.
- When the application codes are uploaded to Google’s servers, the application may behave differently from running on local machines. The reason for this is that the Google’s server has more restrictions, such as 30 seconds limit per request, 1 MB limit per object cached, etc. As there is no way to do debug on Google’s server, these discrepancies are very annoying.
- Querying data from the data store is extremely slow on Google’s server. As a result, the application is intolerably slow if we do not use Memcache to store often used data.
4.1.3 Suggestions to using GAE
Here we would like to provide some suggestions to future teams that may consider using GAE for their web mining project.

- Make the data query as simple as possible. Key based object retrieval is most efficient with GAE. Using sophisticated SQL-style query will often result in errors.
- If a web page need to load multiple types of data, they should split the page request to several phases and store intermediate data in Memcache, or the request process time will most likely exceed the 30 seconds limit.
- For doing data mining, we suggest exporting all the data from GAE to locally installed RDBMS, like SQL Server. Traditional relational DBMS fits better for data mining task of the required scale in this class.

4.2 Data Collection
The data of our project are collected from 4 websites, which are described by source as below.

4.2.1 Gazzetta dello Sport
We collected mainly club and player information from this website (http://www.gazzetta.it). Our spider follows a pattern and goes to each club’s page, which contains the club information and a link to the club squad page. The spider parses the club info and retrieves the contents in the squad page, which are a list of URLs to each player’s profile page. The spider collects the URLs of all the player profile pages and wraps them into task queues for future processing. When the task for parsing one player profile page is processed, a new spider is initialized to go into that player profile page and extract all the player information.

Additionally, we created a spider that collects probable formation for each club. The spider goes to a certain URL for each club, which contains the probable formation in XML format.Parsed data are then stored.

Gazzetta dello Sport provides news feeds of latest Serie A news for users to subscribe. We built an RSS feed parser to collect those news.

4.2.2 BetExplorer
BetExplorer (http://www.betexplorer.com) contains the betting history and match results of all matches since season 1998~1999. Our spider goes to each season’s page and extracts all the match results and corresponding odds.

4.2.3 Football-Italia
Football-Italia (http://football-italia.net/) provides detailed match reports since season 2004~2005. Our spider first goes to a page listing all 38 weeks of this season (2009~2010). On going into the page for each week, the spider collects the URLs for web pages to match reports. We reassemble the URLs so as to collect URLs to matches between the same two teams in other seasons (from season 2004~2005 to season 2008~2009). Hereby we collected match reports for all the relevant matches through seasons. Scores, goal scorers, line-ups and booking information are also collected alongside the match reports.

4.3 API Integration
To improve user experience, we incorporated several open web APIs.
YouTube

YouTube video URLs are collected for each match in this season. When using the YouTube API, we specified the season number, week number, home/away team name and home/away team scores as search criteria for retrieving the videos. The results are ranked by relevance and the first item on result list is stored for each match. By this way, we are able to provide satisfactory video highlight for each match.

Flickr

flickr is used to incorporate pictures for players so that fans can see the player’s pictures in additional to the profile image.

Google Translation

We used a Java-based API (http://code.google.com/p/google-api-translate-java/) that bridges Google Translation and Java. When parsing the Gazzetta dello Sport news feeds, we call this API to directly translate all the news captions and summaries from Italian into English, since our targeted users are all English-readers.

Google Visualization API

Google Visualization API is used in the project to provide users with a ranking trend graph of two opponent teams. By looking at the trend in the graph, people can easily tell whether a team is doing good or bad in either short term or long term.

5. Data Mining Component

The data mining component of the website aims in providing prediction result of a match.

5.1 Preparation for data mining

The training dataset generated for data mining consists of 3484 instances, each of which has 24 attributes. 23 of them are used as input variables and the attribute result is output variable, with 3 different class labels which are homewin, draw and awaywin.

The attributes of one instance cover both long term and short term performances of the two teams (Table 1).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Feature</th>
<th>Description</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>week</td>
<td>Time when the match took place</td>
<td>Home Points</td>
<td>Points earned in the last 4 weeks (home team)</td>
<td>Home Points</td>
<td>Points earned in the last 10 weeks (home team)</td>
</tr>
<tr>
<td>Home Points</td>
<td>Points earned in the last 30 weeks (home team)</td>
<td>Away Points</td>
<td>Points earned in the last 4 weeks (away team)</td>
<td>Away Points</td>
<td>Points earned in the last 10 weeks (away team)</td>
</tr>
<tr>
<td>30week</td>
<td>4week</td>
<td>Homeaway winratio</td>
<td>Home Team Ratio of Home wins to Away Wins</td>
<td>Awayteam homeaway winratio</td>
<td>Away Team Ratio of Home wins to Away Wins</td>
</tr>
<tr>
<td>Away Points</td>
<td>Points earned in the last 30 weeks (away team)</td>
<td>Hometeam Ratio</td>
<td>Home Team Ratio of Draws to Total Matches</td>
<td>Homeavg</td>
<td>Home Team Average goals For in last 10 matches</td>
</tr>
<tr>
<td>30week</td>
<td>4week</td>
<td>of Draws to Total Matches</td>
<td>Away Team Ratio of Draws to Total Matches</td>
<td>Awayavg</td>
<td></td>
</tr>
<tr>
<td>drawratio</td>
<td>Home Team Ratio of Draws to Total Matches</td>
<td>Awaydraw ratio</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homeavg</td>
<td>Home Team</td>
<td>Awayavg</td>
<td>Away Team</td>
<td>Awayavg</td>
<td>Away Team Average</td>
</tr>
</tbody>
</table>
Since the betting companies provide odd for each outcome of a match, it makes more sense to provide probability for each outcome when doing the recommendation. After trying several different methods, developers find out that Logistic Regression can provide the probability distribution for each class label. So the Logistic Regression will be used to do the prediction analysis.

### 5.2 Data mining using Weka

Weka is used to perform the data mining task. The ten coming matches are added in front of the dataset and then the whole dataset is used as training data. In Weka, choose Logistic as classifier and choose to output prediction in more options and the prediction result is obtained (Figure 16).

```plaintext
### Predictions on training set ###

<table>
<thead>
<tr>
<th>inst#, actual, predicted, error, probability distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>
```

**Figure 16 Prediction result in Weka**

### 5.3 Evaluation of the Prediction Model

The accuracy provided by Weka is about 85% (Figure 17).
To evaluate our prediction model, developers set average odds of betting companies as benchmark. Developers calculate the average odds of different betting companies and convert the odds to probabilities of different outcomes. Then prediction results are compared to benchmark (Table 2).

<table>
<thead>
<tr>
<th>Match</th>
<th>hwin</th>
<th>draw</th>
<th>awin</th>
<th>hwin</th>
<th>draw</th>
<th>awin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Napoli - Atalanta</td>
<td>0.65</td>
<td>0.21</td>
<td>0.13</td>
<td>0.546</td>
<td>0.298</td>
<td>0.156</td>
</tr>
<tr>
<td>Udinese - Bari</td>
<td>0.52</td>
<td>0.30</td>
<td>0.17</td>
<td>0.599</td>
<td>0.264</td>
<td>0.138</td>
</tr>
<tr>
<td>AS Roma - Cagliari</td>
<td>0.82</td>
<td>0.12</td>
<td>0.05</td>
<td>0.692</td>
<td>0.219</td>
<td>0.089</td>
</tr>
<tr>
<td>Inter - Chievo</td>
<td>0.84</td>
<td>0.10</td>
<td>0.04</td>
<td>0.684</td>
<td>0.236</td>
<td>0.08</td>
</tr>
<tr>
<td>Juventus - Parma</td>
<td>0.70</td>
<td>0.19</td>
<td>0.10</td>
<td>0.575</td>
<td>0.277</td>
<td>0.148</td>
</tr>
<tr>
<td>Palermo - Sampdoria</td>
<td>0.44</td>
<td>0.28</td>
<td>0.26</td>
<td>0.451</td>
<td>0.343</td>
<td>0.206</td>
</tr>
<tr>
<td>Genoa - AC Milan</td>
<td>0.45</td>
<td>0.28</td>
<td>0.25</td>
<td>0.374</td>
<td>0.336</td>
<td>0.29</td>
</tr>
<tr>
<td>Bologna - Catania</td>
<td>0.38</td>
<td>0.52</td>
<td>0.08</td>
<td>0.438</td>
<td>0.316</td>
<td>0.246</td>
</tr>
<tr>
<td>Livorno - Lazio</td>
<td>0.14</td>
<td>0.25</td>
<td>0.60</td>
<td>0.309</td>
<td>0.381</td>
<td>0.31</td>
</tr>
<tr>
<td>Fiorentina - Siena</td>
<td>0.68</td>
<td>0.21</td>
<td>0.10</td>
<td>0.37</td>
<td>0.415</td>
<td>0.215</td>
</tr>
</tbody>
</table>

Table 2: Comparison between prediction results and benchmark

The table shows clearly that most of the prediction results are quite similar as the average odds.

### 6. Conclusion

Betsmart provides soccer lovers a place where they can get whatever information they want. Also, it provides soccer betters a place to get guidance on how to bet and where to bet. For the future, there are several aspects that Betsmart can go ahead and improve:

1. The scale can be enlarged as it is only focused on Italian Serie A now.

2. The business model can be more versatile, e.g. develop a shopping section and bring in the online shop such as Amazon.

3. Establish the global relationship with betting companies even soccer clubs to make it a real playground for soccer enthusiasts.
7. Appendix

**Team Contribution**

<table>
<thead>
<tr>
<th>Team Member</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ying Jin</td>
<td>Web Design, Data Mining, API Integration</td>
</tr>
<tr>
<td>Cai Chen</td>
<td>Project Manager, API Integration, Data Mining</td>
</tr>
<tr>
<td>Ximing Yu</td>
<td>Architecting, Spidering, Data Pre-processing</td>
</tr>
</tbody>
</table>