Scrubbing Dirty Event Data with an Artificial Neural Net

By: Serena Hillman

School of Interactive Arts and Technology, Simon Fraser University
10153 King George Highway, Surrey, BC V3T 2W1 Canada
shillman@sfu.ca

Abstract

In this paper I will present my case study of an artificial neural network system designed to identify duplicate XML content in live event listings. The application was built to identify duplication of event data including times, dates, venues, event names and cities.

This case study is a niche example of a solution to a large scale dirty data problem. Contributing to advancements in data scrubbing concepts is important because of the explosion of content feeds, user-generated content and mash-up applications that leave site aggregate content, such as comparison shopping engines and affiliate sites with dirty data. My hypothesis is that by applying heuristics, fuzzy logic, and an artificial neural network, the software can identify similar data.

Keywords
Mashup, Data Scrubbing, XML, Duplicate Content, Event Listings, Matching, Dirty Data, India UID Project

Introduction

For the case study the application will be applied to ShowTimeTickets.com's event data. ShowTimeTickets.com is an international ticket broker located in Vancouver, BC and sells tickets to events worldwide.

The company's unique selling proposition is focused around having the most tickets available anywhere. At the time of testing the ShowTimeTickets.com database housed over forty-five thousand event listings.

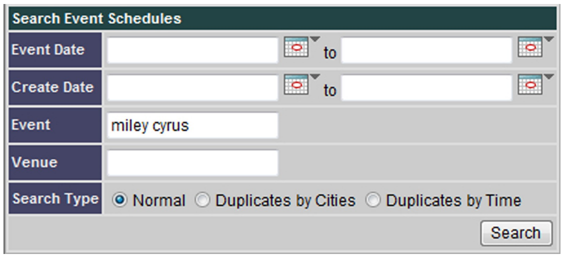
In order to obtain the most tickets online ShowTimeTickets.com receives event data instantly from five different sources: Razorgator, Stubhub, TicketNetwork, Pollstar and Event Inventory. All five of these sources provide their own specific inventory and event listings. Aggregating this data provides a unique advantage in the the marketplace. The feeds have varying levels of data quality; some are from larger ticket exchange sites with high quality control and others are not. With the aggregation of these XML feeds comes heaps of dirty data in the form of duplicate event listings.

When data is imported to the ShowTimeTickets.com database it is done so independently. For example, venue, cities/countries, events/event schedules, each have a unique identification number. Each feed is scheduled to import once per day.

When an import occurs ticket data is not captured at this time. Instead, the URL is captured and pinged in real time when a request from ShowTimeTickets.com for the particular event occurs. This way ticket data is never stale.

Background

Prior to implementation of the Artificial Neural Net ShowTimeTickets.com has taken a more manual approach to identifying and fixing duplicates. The tools created for identifying and merging are housed within the website's content management system. Through use of these tools a staff member can search newly imported data and compare the new events to the old events (see figure 1). This exercise is done daily and on average takes roughly three hours to complete.



Search Event Schedules

Event: miley cyrus

Search Type: ● Normal ○ Duplicates by Cities ○ Duplicates by Time

Figure 1 – ShowTimeTickets.com's CMS Merging Tool

This daily task involves a ShowTimeTickets.com staff member to search for all newly imported events for the day and compares each, one-by-one, to the events already within the database. Before the staff member sees the data as newly imported, some general rules are applied. These rules include the following:

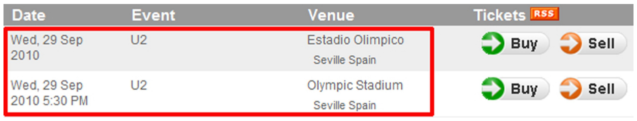
If the new data is an exact match, or the data is the same when removing “the,” spaces and /or special characters--merge with the current data.

(These rules are applied to every piece of data separately, including: venue, location, event name, etc.).

If data has been merged at a prior date, the system will not re-import the data unless the data has changed. If the data has changed then the data gets updated. For example, if the event “the Rolling Stones” and another event “Rolling Stone” are merged together and one of the events for the “Rolling Stones” gets updated, then the merged event will update automatically.

Another way ShowTimeTickets.com identifies duplicates is through a feedback form on the website. A link to the form is positioned next to every event's event dates with the anchor text “Report a Listing Error.” When the user clicks on the link they are presented with an optional text field asking what type of issue they see. On average ShowTimeTickets.com receives about one event issue per day.

While in theory following all these standard operating procedures should eliminate all duplicates it is clear that human error on such a repetitive and mundane task results in a high error rate. Later on when we review the results of the artificial neural net we see the proof of just how ineffective the current set-up has been.

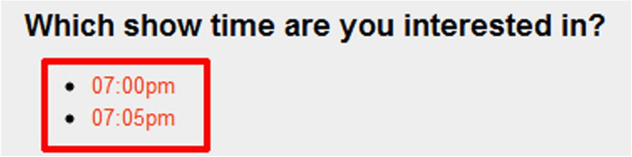


Date	Event	Venue	Tickets RSS
Wed, 29 Sep 2010	U2	Estadio Olimpico Seville Spain	Buy Sell
Wed, 29 Sep 2010 5:30 PM	U2	Olympic Stadium Seville Spain	Buy Sell

Figure 2 – an example of ShowTimeTickets.com Duplicates

Figure 2 shows a screen shot of a normal duplication issue on ShowTimeTickets.com. With this example we see duplication for two reasons. One of the event listings is from Pollstar, and this particular source does not supply times of events, while the other event is supplied by Event Inventory which does supply the times for the event. Duplication occurs because the events have different times. The second reason for content duplication is the naming of the venue. One event has the American name for the venue while the other has the Spanish name.

ShowTimeTickets.com is not the only website experiencing these content duplication issues.

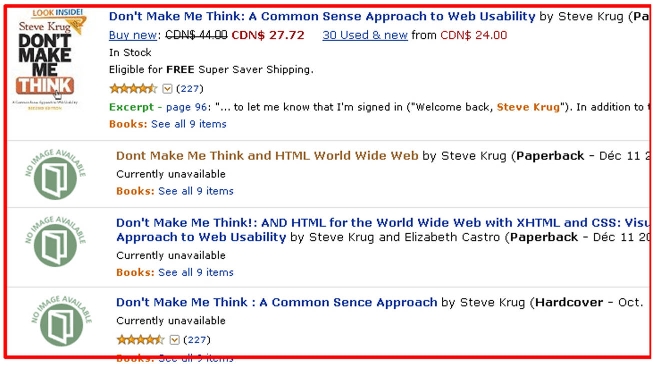


Which show time are you interested in?

- 07:00pm
- 07:05pm

Figure 3 – TicketFlow.com listing for Vancouver Canucks vs. Edmonton Oilers November 28, 2009

Figure 3 shows a screen shot of a typical aggregation comparison website. In this example you can see two event times of 7:00pm and 7:05pm being offered as different events for a Vancouver Canucks game. Not only will the user be confused as to which game is correct, but both listings will have different ticketing information making a consumer's ability to compare all available tickets difficult.



Don't Make Me Think: A Common Sense Approach to Web Usability by Steve Krug (Pa

Buy new: ~~CDN$ 44.00~~ CDN$ 27.72 30 Used & new from CDN$ 24.00

In Stock

Eligible for FREE Super Saver Shipping.

★★★★☆ (227)

Excerpt - page 96: "... to let me know that I'm signed in ("Welcome back, Steve Krug"). In addition to

Books: See all 9 items

Dont Make Me Think and HTML World Wide Web by Steve Krug (Paperback - Déc 11 2

Currently unavailable

Books: See all 9 items

Don't Make Me Think!: AND HTML for the World Wide Web with XHTML and CSS: Visu Approach to Web Usability by Steve Krug and Elizabeth Castro (Paperback - Déc 11 20

Currently unavailable

Books: See all 9 items

Don't Make Me Think : A Common Sence Approach by Steve Krug (Hardcover - Oct.

Currently unavailable

★★★★☆ (227)

Books: See all 9 items

Figure 4 - Amazon Search for “Don't Make Me Think”

Figure 4 shows the Amazon marketplace results for a popular web usability book, Steve Krug's *Don't Make Me Think*. As we see in Figure 4, Amazon returns multiple results for the same book, with all these listings having different ratings and descriptions. One can guess that the duplication on Amazon is due to user-generated content.

According to a study done in 2006 by Neilsen NetRatings, a global leader in internet media and market research, user-generated content drives half of the United States' top ten fastest growing brands.

Table 1.

Brand	Jul'05	Jul'06	% Growth
HSBC	1290	6377	394.00%
Sonic Solutions	1098	3740	241.00%
Associated Press	2901	9692	234.00%
ImageShack	2324	7745	233.00%
Heavy.com	965	3021	213.00%
Flickr	2105	6346	201.00%
ARTIST Direct	1131	3219	185.00%
Partypker.com	2127	6043	184.00%

MySpace	16239	46025	183.00%
Wikipedia	10387	29176	181.00%

*fastest growing web brands among those with a minimum unique audience of 750,000 in July 2006 (Bausch and Han 1988).

With the user-generated information statistics reported above, duplication and classification is a real concern for the future of the web.

System Overview

The ShowTimeTickets.com development team was in charge of the development of the neural net while project management and testing was done by myself.

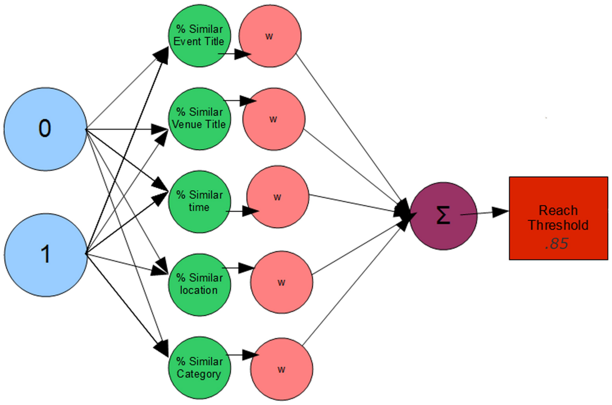


Figure 5 – Artificial neural net for scrubbing dirty event data

The neural net (figure 5) starts with two events. Event 0 is the newly imported event and event 1 is an event already in the ShowTimeTickets.com database.

These events are then compared to each other on five levels of similarity. These levels include each piece of data obtained for the source: event, event date, location, category and venue.

To determine similarity we reviewed many fuzzy string open source SIM Metrics libraries. To determine which one was best for our situation we tested them against data from the ShowTimeTickets.com production database.

These tests resulted in the following findings:
Findings - SIM Metrics Libraies Results

After reviewing the results it was determined that the JaroWinkler Library produced the best results for our particular issues.

The Jaro-Winkler distance is a measure of similarity between two strings (Winkler, 1999). It is a variant of the Jaro distance metric (Jaro, 1989, 1995) and mainly used in the area of record linkage (duplicate detection). The higher the Jaro-Winkler distance for two strings is, the more similar the strings are. The Jaro-Winkler distance metric is designed and best suited for short strings such as person names. The score is normalized such that 0 equates to no similarity and 1 is an exact match.

The Jaro distance metric states that given two strings s_1 and s_2, their distance d_j is:

$$d_j = \frac{1}{3}\left(\frac{m}{|s_1|}+\frac{m}{|s_2|}+\frac{m-t}{m}\right)$$

(Wikipedia contributors, 2009)

Defining category similarity as the only exception to not using the JaroWinkler library. Instead similarity was determined by utilizing the current category mapping tool.

The category mapping tool is located in the Content Management System for the ShowTimeTickets.com website. This mapping tool contains all the categories for the sources as well as ShowTimeTickets.com categories (see figure 6). A ShowTimeTickets.com staff member will map the appropriate categories together. Similarity for categories is based on this information by assigning a similarity percentage to the distance of the the two categories.

After every node has established a similarity, % weights are then applied to the results depending on how important similar data in the area is in identifying duplicates.

Results

The first partial run on the production database had weights set to the following:
weightSimilarEventTitle = 0.25F;
weightSimilarVenueTitle = 0.1F;
weightSimilarTime = 0.3F;
weightSimilarCity = 0.1F;
weightSimilarState = 0.1F;
weightSimilarCountry = 0.1F;
weightSimilarEventCategory = 0.05F;

These weights gave us these results:
Artificial Neural Network Results 1

The two red highlighted items are false positives. In order to counter the false positives we lowered the weights for the event title similarity and increased the weights for venue similarity.

This put the weights at:
weightSimilarEventTitle = 0.20F
weightSimilarVenueTitle = 0.15F
weightSimilarTime = 0.3F
weightSimilarCity = 0.1F
weightSimilarState = 0.1F
weightSimilarCountry = 0.1F
weightSimilarEventCategory = 0.05F;

Finally, the sum weighted results gave us a final output. This output was weighed against a threshold of .85, which we found through multiple tests of thresholds, .85 was indeed the best.

Initial tests on December 18th, 2009 identified 5,965 duplicate events running on the ShowTimeTickets.com production environment. Of these results, less than 1% were identified as false positives. Results can be viewed here: Artificial Neural Network Results 2

The first run of the application would have taken an estimated forty-six hours. After optimizing the code and putting the application in the same network as the database we were able to get run-time down to nine hours. The first full run compared the entire database of more then forty-five thousand events.

Implementation

The application will be set to run weekly and produce a report in csv format similar to the Artificial Neural Network Results 2.

After the report had identified duplicates in the system, a ShowTimeTickets.com staff member will merge the duplicate events. If events are identified but no merge action is required then the identification number is placed into an exception application that will remove it from future reports.

An example of when no action would be required would be if a band is opening for another band and thereby has the same venue, location and time event data. In this case, having both events listed on our site is ideal, especially since some bands co-tour.

Findings

The results so far have been quite positive, after tweaking the weights based on the initial results and playing around with the threshold a 99% success rate is hopeful.

Besides fixing the initial problem of finding duplicates, the duplicate report has also allowed us to identify bands that are touring together. This is important because the customer management system allows us to link these events together and allows for the sharing of tickets because the event is the same.

Overall the report provides ShowTimeTickets.com with some much needed statistics on the sources which events are imported from, and ultimately more information on how to fix the duplication problems ShowTimeTickets.com face on a daily basis.

Future Research

The problem of duplication on ShowTimeTickets.com's website is really a symptom of overall content organization online. With such a large scale issue more regulatory practices are needed from online players with authority. How can companies such as Microsoft, Google and Yahoo! combat this issue.

Google's mission statement clearly identifies fixing these issues as a main priority; it reads “to organize the world's information and make it universally accessible and useful” (Google, 1998).

In November of 2009 when the India government announced the UID (Unique Identification) Project all three companies were quick to offer assistance (Domain-b, 2009). Perhaps these major companies are looking at the development of India's UID project as a blueprint for creating unique identification for all things online.

References

Bausch S., Han L. 2006. *User-Generated Content Drives Half of U.S. Top 10 Fastest Growing Web Brands, According to Nielsen//NetRatings.* Nielsen//NetRatings. http://www.nielsen-online.com/pr/PR_060810.PDF

Chapman, S. 2006. *SimMetrics*, Natural Language Processing Group, Dept. O Computer Science, Univ. Sheffield, Sheffield, UK. http://www.dcs.shef.ac.uk/~sam/simmetrics.html

Domain-b. 2009. Google, Yahoo, Intel, Microsoft, IBM and Cisco vie for UID project news. http://www.domain-b.com/infotech/itnews/20091112_uid_project_oneView.html

Google. 1998. *Company Overview*, Corporate Information. http://www.google.com/corporate/

Jaro, M. A. (1989). *Advances in record linking* *methodology as applied to the 1985 census of Tampa Florida.* Journal of the American Statistical Society **84** (406): 414–20.

Jaro, M. A. (1995). *Probabilistic linkage of large public health data file*. Statistics in Medicine **14** (5-7): 491–8.

McCarthy, M. 2007. *Nielsen/NetRatings' August social media numbers: Not much change,* CNet News, The Social. http://news.cnet.com/8301-13577_3-9777942-36.html

Wikipedia Contributes. 2009. *Jaro-Winkler distance, Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Jaro-Winkler_distance&oldid=331390384

Winkler, W. E. (1999). *The state of record linkage and current research problems*. Statistics of Income Division, Internal Revenue Service Publication R99/04.

Winkler, W. E. (2006). *Overview of Record Linkage and Current Research Directions*. Research Report Series, RRS.