Reverse Engineering Google Ranking Algorithm

Powered By
Netezza, Fuzzy Logix, & Virante
Problem Statement

We would like businesses to be able to make effective, efficient decisions in determining how and where to allocate resources to improve their rankings in Google.

Today, because Google neither reveals its ranking algorithm nor its data set, businesses are likely to waste time and resources using experimental, untrustworthy, and in many cases risky methods to improve their rankings. Businesses must rely on experience and anecdotal evidence from search agencies to determine the efficacy of these methods. Mistakes can be very costly to businesses: from unrealized returns on investment to public affairs nightmares such as those recently experienced by J.C. Penney and Overstock.com.

We intend to determine the factors and their weights behind Google's ranking algorithm by applying Fuzzy Logix's proprietary in-database computation on IBM Neteeza to data acquired from Google, search data vendors and the web at large.
About Trident Marketing

At Trident Marketing, our people are our greatest asset, and it’s been that way for over 20 years. Founded in 1986 with the simple proposition - to envision, to achieve and to enjoy - the original team of a few young entrepreneurs laid the foundation for what has become a highly successful process of direct response sales. With such success, Trident Marketing is recognized as one of the area’s most respected and profitable organizations. Trident Marketing is one of Moore County, N.C.’s largest employers. With over 400 associates, most of whom are located in its Southern Pines headquarters, we enjoy producing national results while staying locally focused.
About Fuzzy Logix

• Founded based on 12+ years of research by the partners in areas of quantitative and high performance computational methods

• Ported and released the first independent library of in-database analytics on Netezza, followed by Sybase IQ, SQL Server, Informix, ParAccel, mySQL and Asterdata.

• The only multi-platform library of in-database models in the world

• Many Fortune 1000 customers, reseller and OEM partners

• Over 800 analytic functions with extensive financial models available

• Released the first general library of analytics that leverage graphic processing units (GPUs) via .net and .dll

• World class team with vast experience in analytics

• Consulting and training solutions that include custom solution development
About Virante, Inc.

- Located in Research Triangle Park (NC)
- Provide Search Engine Marketing (Organic SEO and Paid Search services)
- Data and Research-Driven search marketing solutions
- Ethical SEO strategies with white-glove service

Virante has been a part of the web marketing industry since the late 1990's. Focused specifically on Search Engine Optimization in its early years, Virante earned a reputation of "zero to one million", empowering numerous companies to grow to a million in sales. Virante's expansion into the remainder of online marketing has been similarly fruitful, creating the "rent-to-own" strategy for building and sustaining traffic and qualified leads for any web site or business. Today, Virante offers a research and data-driven approach to providing white-glove search engine marketing solutions for their clients.

Learn more about our research at Virante.com and TheGoogleCache.com
Reverse Engineering Google’s Ranking Algorithm

• **Offsite Factors (Importance and Relevance):**
  - Google's Primary Contribution to Search Engines was the "Citation-Based Weighting".
  - Link Quality Measures
    - PageRank (MozRank, ACRank)
    - Trust Rank (MozTrust)
  - Link Quantity Measures
    - Internal
    - External, Unique Domains, Unique IPs, Unique Class-C
  - Link Relevance Measures
    - Anchor Text (not included in study)
• **Offsite Data Sources**
  – SEOMoz Site Intelligence API (www.seomoz.org)
  – Majestic SEO Enterprise API (www.majesticseo.com)
Reverse Engineering Google’s Ranking Algorithm

• **Onsite Keyword Usage and URL Factors**
  – Google Still Relies on On-Site and On-Page Factors for Relevancy
  – Keyword usage in HTML (Title, Meta Description, Hx Tags)
  – Keyword usage in URL, Exact-Match Domain

• **Onsite Data Sources**
  – Keyword Data: KeywordDiscovery ([www.keyworddiscovery.com](http://www.keyworddiscovery.com))
  – Google Rankings: AuthorityLabs ([www.authoritylabs.com](http://www.authoritylabs.com))
  – Page Data: Virante ([www.virante.com](http://www.virante.com))

• **Type of Data** - Competitive Search Terms
Reverse Engineering Google’s Ranking Algorithm

- **Data Collection and Filtering**
  - Keyword Scrubbing:
  - Remove Local Search
  - Remove Non-Competitive Keywords

- **SERP Scrubbing (Search Engines Results Page)**
  - Remove Double Listings and Branded Search
  - Remove Google and YouTube Inserts
  - Remove Local Search listings
  - Limit to top 10
Reverse Engineering Google’s Ranking Algorithm

• **Modeling Process**
  – Data for 1 million searches, 1.4 million domains, 5.6 million URLs
  – Generally search terms have ranks 1..10, some have less ranks
  – *Statistical Technique – Multinomial Logistic Regression*
  – Collapsing the ranks in 3 levels
    - Best rank – 1
    - Top ranks - 2, 3, 4
    - Mid tier – 5, 6, 7
    - Low ranks – 8, 9, 10
  – Modeling done with 10-20% of the data, 80-90% retained as holdout
  – Sampling performed repeatedly to test for stability of coefficients
Reverse Engineering Google’s Ranking Algorithm

• Single Factor Analysis
  – Analysis of individual factors for predictive power
  – Statistical analysis shows that a lot of factors do not have any predictive power – contradicts popular belief
    • Presence of keyword in title, description, headers, body, etc.
    • Presence of the keyword in URL
    • Keyword in bold or italics
    • Keyword density on the page
    • Exact match (yes/no) of the keyword with domain
  – Yet, people talk about these factors all the time
Multi Factor Modeling

Derivation of composite factors

- Composite metric based on Folders and URL length – large number of folders and very long URLs cause rankings to drop
- Composite ranking based on quality of inbound links and outbound links to the domain and URL – data from SEOMoz
- Jaro Winkler Score of the keyword in the domain and URL
- Ratio of Jaro Winkler Score for domain and URL
- Jaro Winkler Scoring – algorithm to compute a match between two words

Composite factors eliminate the issue of multi-collinearity and yet preserve all the factors important for this exercise
Reverse Engineering Google’s Ranking Algorithm

• Model Performance
  – Gini Coefficient ~ 40%, 40% may not be a great number but for this exercise it is quite significant

<table>
<thead>
<tr>
<th>Google Rank</th>
<th>Our Rank</th>
<th>Number of Occurrences</th>
<th>% of Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>240419</td>
<td>26.41%</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>155127</td>
<td>17.04%</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>118447</td>
<td>13.01%</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>95894</td>
<td>10.53%</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>81150</td>
<td>8.91%</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>68883</td>
<td>7.57%</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>57783</td>
<td>6.35%</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>46207</td>
<td>5.08%</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>32206</td>
<td>3.54%</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>14303</td>
<td>1.57%</td>
</tr>
</tbody>
</table>

• Model is much better than random
• First three ranks – lift from the model is 26%
• We are not expected to match exactly with Google
### Reverse Engineering Google’s Ranking Algorithm

<table>
<thead>
<tr>
<th>Search Term</th>
<th>URL</th>
<th>Google Rank</th>
<th>Our Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>green bay packers schedule</td>
<td><a href="http://www.nfl.com/teams/schedule?team=GB">http://www.nfl.com/teams/schedule?team=GB</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers news</td>
<td><a href="http://packersnews.greenbaypressgazette.com/">http://packersnews.greenbaypressgazette.com/</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers logo</td>
<td><a href="http://en.wikipedia.org/wiki/Green_Bay_Packers">http://en.wikipedia.org/wiki/Green_Bay_Packers</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers apparel</td>
<td><a href="http://www.footballfanatics.com/NFL_Green_Bay_Pack">http://www.footballfanatics.com/NFL_Green_Bay_Pack</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers graham harrell</td>
<td><a href="http://www.packers.com/team/roster/Graham-Harrell/">http://www.packers.com/team/roster/Graham-Harrell/</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers roster</td>
<td><a href="http://www.packers.com/team/players.html">http://www.packers.com/team/players.html</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers pictures</td>
<td><a href="http://www.picssearch.com/pictures/teams%20and%20sp">http://www.picssearch.com/pictures/teams%20and%20sp</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers pro shop</td>
<td><a href="http://www.packersproshop.com/">http://www.packersproshop.com/</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packer tickets</td>
<td><a href="http://www.packers.com/tickets/index.html">http://www.packers.com/tickets/index.html</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers clothing</td>
<td><a href="http://www.footballfanatics.com/NFL_Green_Bay_Pack">http://www.footballfanatics.com/NFL_Green_Bay_Pack</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers logo pictures</td>
<td><a href="http://www.sportslogos.net/team.php?id=171">http://www.sportslogos.net/team.php?id=171</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers tickets</td>
<td><a href="http://www.packers.com/tickets/index.html">http://www.packers.com/tickets/index.html</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers zebra cowboy hat</td>
<td><a href="http://www.amazon.com/Cheese-Cowboy-Hat-Green-Pack">http://www.amazon.com/Cheese-Cowboy-Hat-Green-Pack</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packer apparel</td>
<td><a href="http://www.footballfanatics.com/NFL_Green_Bay_Pack">http://www.footballfanatics.com/NFL_Green_Bay_Pack</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packer earrings</td>
<td><a href="http://www.amazon.com/Cheese-Earrings-Green-Bay-Pa">http://www.amazon.com/Cheese-Earrings-Green-Bay-Pa</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers desktop background</td>
<td><a href="http://www.packerscentral.com/desktops.php">http://www.packerscentral.com/desktops.php</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packer scrubs</td>
<td><a href="http://www.nflshop.com/family/index.jsp?categoryid">http://www.nflshop.com/family/index.jsp?categoryid</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers wallpaper</td>
<td><a href="http://www.packers.com/fan-zone/wallpaper.html">http://www.packers.com/fan-zone/wallpaper.html</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packer tshirts</td>
<td><a href="http://www.footballfanatics.com/NFL_Green_Bay_Pack">http://www.footballfanatics.com/NFL_Green_Bay_Pack</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers background</td>
<td><a href="http://www.packers.com/fan-zone/wallpaper.html">http://www.packers.com/fan-zone/wallpaper.html</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers earrings</td>
<td><a href="http://www.amazon.com/Cheese-Earrings-Green-Bay-Pa">http://www.amazon.com/Cheese-Earrings-Green-Bay-Pa</a></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>green bay packers</td>
<td><a href="http://www.packers.com/">http://www.packers.com/</a></td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
• **The Netezza Advantage**
  
  – Very large amount of data
  
  – We can mine the factors individually and build the model in about 5 minutes using Netezza TwinFin and the in-database analytics library DB Lytix from Fuzzy Logix
  
  – We collect data on a daily basis and check for deviations, monitor model performance and mine search results for additional factors
  
  – *We can mine millions of search results without worrying about the load on the system and hampering production systems.*
  
  – *This modeling is done during peak hours on a system running production reports and tracking sales.*
Reverse Engineering Google’s Ranking Algorithm

• **Future Initiatives**
  – Additional Factors
  – Latent Dirichlet Allocation
  – Anchor Text Statistics
  – Duplicate Content
  – Data Collection and Filtering
  – Rapid Data Acquisition
  – Paid Search
Reverse Engineering Google’s Ranking Algorithm

Why is this important for your organization?

Case Study - Trident Marketing

- High Volume Customer Acquisition Opportunities ( Allows us to compete)
- Low Cost (Compared to traditional marketing channels, SEO vs PPC)
- Better Customer (Higher Credit, Less Churn, etc)
- Differentiation From Many Performance Based Marketing Companies (Many do not offer)
- Allows For Easier Market Entry (Low cost, in-house processes)
- Efficiency (Knowing factors, develop systems around)
- Scalable (Ability to grow multiple organic programs under one roof)