Query Rewriting through Link Analysis of Click Graph

Alekh Jindal
Motivation

Today's Most Profitable Investments
23 Mar 2006 ... In this issue, Dr. Mark Skousen shares his two most profitable investments and wealth-building strategies. His answers may surprise you.
www.investmentu.com/UEL/2006/20060323.html - 37k - Cached · Similar pages

Agypten issue tracker: Issue 1086: MONEY AND GOLD DUST FOR ...
Title: MONEY AND GOLD DUST FOR PROFITABLE INVESTMENT. Priority, Status, unread. Superseder, Nossy List, troshuv.troshuv, troshuv.troshuv1 ...
www.inetration.de/roundup/aegypten/issue1086 - 13k - Cached · Similar pages

Pakistan- A Profitable Investment Destination
Pakistan Embassy Berlin: Visa application form, Passport application form.
www.pakemb.de/index.php?id=126 - 10k - Cached · Similar pages

dict.cc dictionary :: profitable investment · English-German ...
dict.cc English-German Dictionary: Translation for profitable investment.
www.dict.cc/english-german/profitable+investment.html - 9k - Cached · Similar pages

Profitable Investment Portfolio
A number of diplomats and Ambassadors have expressed the keen desire of their respective governments to explore opportunities of profitable investments in .......
www.profitableinvestmentportfolio.com/ - 83k - Cached · Similar pages

Learn how to Invest, Buy Stocks, Sell Stocks, Investing and ...
To bring you in a position to accomplish profitable investments in equities, easily manage to make money now and plan for your financial future and ...
www.greeksahres.com/ - 44k - Cached · Similar pages

Schneider Electric - Training - Your benefits - Training, a ...
Training, a profitable investment. Return on investment: immediate and over time.
Sponsored Search

- Ads relevant to user query shown above or alongside search results
- Each bid has query(q), ad(\(\alpha\)) and price(p)
Sponsored Search

- Ads relevant to user query shown above or alongside search results
- Each bid has query($q$), ad($\alpha$) and price($p$)
Query Rewrites

- Not direct bids for many queries
- Ads have little text; lesser information
- Rewrites: similar queries based on history of ads displayed and clicked
Query Rewrites

- Not direct bids for many queries
- Ads have little text; less information
- Rewrites: similar queries based on history of ads displayed and clicked

Diagram:
- Front End: q \(\rightarrow\) q, rewrites for q \(\rightarrow\) Back End
- Back End: History \(\rightarrow\) Sponsored search results
- Ads \(\rightarrow\) Bids
Click Graph
Click Graph

- Generated by back end
- Directed, weighted, bipartite graph
- Formally: $G = (Q, A, E)$
  - $Q$: set of queries $q$
  - $A$: set of ads $\alpha$
  - $E$: set of edges $e$ from $q$ to $\alpha$, s.t. at least one user that issued $q$ clicked on $\alpha$
- Edge weights:
  - Impressions
  - Clicks
  - Expected click rate
Click Graph - example

Query
- pc
- camera
- Digital camera
- tv
- flower

Ad
- hp.com
- bestbuy.com
- teleflora.com
- orchids.com
Goal: find similar queries
Intuition: queries with common ad clicks are similar
Analogous to collaborative filtering (CF)
• Users as queries; Recommendation as ads
Query Similarity
Query Similarity

Similar due to inheritance

Similar due to...
uh...other factors
Query Similarity

Guys are similar if they like the similar girl!
Guys are similar if they like the similar girl! ... and vice-versa!
Similarity: naïve approach

Idea: count number of common ads
Similarity: naïve approach

Idea: count number of common ads

Query:
- pc
- camera
- Digital camera
- tv
- flower

Ad:
- hp.com
- bestbuy.com
- teleflora.com
- orchids.com
Similarity: naïve approach

Idea: count number of common ads

<table>
<thead>
<tr>
<th>Query</th>
<th>Ad</th>
<th>pc</th>
<th>camera</th>
<th>digital camera</th>
<th>tv</th>
<th>flower</th>
</tr>
</thead>
<tbody>
<tr>
<td>pc</td>
<td>hp.com</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>camera</td>
<td>bestbuy.com</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Digital camera</td>
<td>teleflora.com</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>tv</td>
<td>teleflora.com</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>flower</td>
<td>orchids.com</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>
**Similarity: naïve approach**

**Idea:** count number of common ads

**Problem:** $\text{sim}(\text{pc, tv}) = 0$

<table>
<thead>
<tr>
<th></th>
<th>pc</th>
<th>camera</th>
<th>digital camera</th>
<th>tv</th>
<th>flower</th>
</tr>
</thead>
<tbody>
<tr>
<td>pc</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>camera</td>
<td>1</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>digital camera</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>tv</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>0</td>
</tr>
<tr>
<td>flower</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>
Similarity: Simrank

Idea: Two objects are similar if they are referenced by similar objects
Idea: Two objects are similar if they are referenced by similar objects.
Idea: Two objects are similar if they are referenced by similar objects
Idea: Two objects are similar if they are referenced by similar objects.
Similarity: Simrank

Idea: Two objects are similar if they are referenced by similar objects

a  ![Diagram](image1.png)

b  ![Diagram](image2.png)
Similarity: Simrank

Idea: Two objects are similar if they are referenced by similar objects
Idea: Two objects of one type are similar if they are referenced by similar objects of second type.
Idea: Two objects of one type are similar if they are referenced by similar objects of second type.
Formally: $E(x)$ is the set of neighbors of $x$

- $N(x)$ is the number of neighbors of $x$

For queries $q$ and $q'$, similarity $s(q, q')$ is given as:

$$s(q, q') = \frac{C_1}{N(q)N(q')} \sum_{i \in E(q)} \sum_{j \in E(q')} s(i, j)$$

Similarly, for ads $\alpha$ and $\alpha'$, similarity $s(\alpha, \alpha')$ is given as:

$$s(\alpha, \alpha') = \frac{C_2}{N(\alpha)N(\alpha')} \sum_{i \in E(\alpha)} \sum_{j \in E(\alpha')} s(i, j)$$

- $C_1, C_2$ are constants in $[0,1]$
Example: Find $s(a,c)$

Let $C_1 = 0.8$

$$s(a,c) = \frac{C_1}{N(a)N(c)} \sum_{i \in E(a)} \sum_{j \in E(c)} s(i,j)$$

$$s(a,c) = 0.8 \cdot s(d,e)$$

Iteration 1:
$s(x,x) = 1$, $s(x,y) = 0$
$s(a,c) = 0$
Similarity: Bipartite Simrank

- Iteration 2:

\[
s(a, c) = 0.8 \cdot s(d, e)
\]

\[
s(a, c) = 0.8 \cdot \frac{C_2}{N(d)N(e)} \sum_{i \in E(d)} \sum_{j \in E(e)} s(i, j)
\]

\[
s(a, c) = 0.8 \cdot \left\{ \frac{0.8}{2 \times 2} \right\}
\]

\[
(s(a, b) + s(a, c) + s(b, b) + s(b, c)) \}
\]

\[
s(a, c) = 0.32 \cdot \{ (0 + 0 + 1 + 0) \}
\]

\[
s(a, c) = 0.32
\]
Idea: Two objects of one type are similar if they are referenced by similar objects of second type

<table>
<thead>
<tr>
<th></th>
<th>pc</th>
<th>camera</th>
<th>digital camera</th>
<th>tv</th>
<th>flower</th>
</tr>
</thead>
<tbody>
<tr>
<td>pc</td>
<td>-</td>
<td>0.619</td>
<td>0.619</td>
<td>0.437</td>
<td>0.000</td>
</tr>
<tr>
<td>camera</td>
<td>0.619</td>
<td>-</td>
<td>0.619</td>
<td>0.619</td>
<td>0.000</td>
</tr>
<tr>
<td>digital camera</td>
<td>0.619</td>
<td>0.619</td>
<td>-</td>
<td>0.619</td>
<td>0.000</td>
</tr>
<tr>
<td>tv</td>
<td>0.437</td>
<td>0.619</td>
<td>0.619</td>
<td>-</td>
<td>0.000</td>
</tr>
<tr>
<td>flower</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>-</td>
</tr>
</tbody>
</table>
Idea: Two objects of one type are similar if they are referenced by similar objects of second type

\[ s(\text{camera}, \text{tv}) = s(\text{camera}, \text{digital camera}) \]
“evidence” not taken into account

\[ \text{sim}(a,b) < \text{sim}(c,d) \]

Expected: \( \text{sim}(a,b) > \text{sim}(c,d) \)
“evidence”: Number of common neighbours

Evidence function:

\[
evidence(a, b) = \sum_{i=1}^{\frac{|E(a) \cap E(b)|}{2^i}}
\]

Revised Simrank:

\[
\begin{align*}
    s_{evidence}(q, q') &= evidence(q, q') \cdot s(q, q') \\
    S_{evidence}(\alpha, \alpha') &= evidence(\alpha, \alpha') \cdot s(\alpha, \alpha')
\end{align*}
\]
Similarity: Evidence Simrank

- \( \text{sim}(a,b) > \text{sim}(c,d) \) after 1\text{st} iteration
Consistency Rules

- If variance is less and edge weight more then similarity is more

Expected: $\text{sim}(a, b) > \text{sim}(a, c)$
Consistency Rules

For equal variance, if edge weight is more then similarity is more

Expected: $\text{sim}(a,b) > \text{sim}(a,c)$
Similarity: Weighted Simrank

- Transition probability:
  \[ p(x, i) = \text{spread}(i) \cdot \text{normalized\_weight}(x, i) = W(x, i) \]
  \[ \text{spread}(i) = e^{-\text{variance}(i)} \]
  \[ \text{normalized\_weight}(\alpha, i) = \frac{W(\alpha, i)}{\sum_{j \in E(\alpha)} W(\alpha, j)} \]

- Revised Simrank:
  \[ s_{\text{weighted}}(q, q') = \text{evidence}(q, q') \cdot C_1 \cdot \sum_{i \in E(q)} \sum_{j \in E(q')} W(q, i) W(q', j) s_{\text{weighted}}(i, j) \]
  \[ s_{\text{weighted}}(\alpha, \alpha') = \text{evidence}(\alpha, \alpha') \cdot C_2 \cdot \sum_{i \in E(\alpha)} \sum_{j \in E(\alpha')} W(\alpha, i) W(\alpha', j) s_{\text{weighted}}(i, j) \]
Scalability

- Query rewrites offline and in batch
- Space required: $O(N^2)$
  - $N$: total number of nodes (query+ad)
- Time Required: $O(kN^3)$
  - $k$: number of iterations
  - typical value, $k=7$
- Time complexity can be reduced to: $O(kN^2d)$
  - $d$: average of $N(a).N(b)$
  - $d$ does not grow with $N$
- For 15 million queries, 14 million ads and 28 million edges, Simrank++ completes in 6 hours on a single machine
Experiments: baselines

- Three query rewriting techniques

  - Pearson:
    \[
    \text{sim}_{\text{pearson}}(q, q') = \frac{\sum_{\alpha \in E(q) \cap E(q')} (w(q, \alpha) - \bar{w}_q) (w(q', \alpha) - \bar{w}_{q'})}{\sqrt{\sum_{\alpha \in E(q) \cap E(q')} (w(q, \alpha) - \bar{w}_q)^2 (w(q', \alpha) - \bar{w}_{q'})^2}}
    \]

  - Jaccard:
    \[
    \text{sim}_{\text{jaccard}}(q, q') = \frac{|E(q) \cap E(q')|}{|E(q) \cup E(q')|}
    \]

  - Cosine:
    \[
    \text{sim}_{\text{cosine}}(q, q') = \arccos \frac{\nu(q) \cdot \nu(q')}{\|\nu(q)\| \|\nu(q')\|}
    \]
Experiments: baselines

- **Example**
  - \( \text{sim}_{\text{pearson}} (Q, Q') = 1.414 \)
  - \( \text{sim}_{\text{Jaccard}} (Q, Q') = 1.000 \)
  - \( \text{sim}_{\text{cosine}} (Q, Q') = 0.841 \)
Experiments: baselines

Example

- $\text{sim}_{\text{pearson}}(Q, Q') = 1.414$
- $\text{sim}_{\text{Jaccard}}(Q, Q') = 1.000$
- $\text{sim}_{\text{cosine}}(Q, Q') = 0.841$

However,

- $\text{sim}_{\text{pearson}}(Q, b) = 0$
- $\text{sim}_{\text{Jaccard}}(Q, b) = 0$
- $\text{sim}_{\text{cosine}}(Q, b) = 0$
Experiments: Dataset

- Two week click graph from US Yahoo! Search
  - 15 million queries, 14 million ads, 28 million edges
- Edge weight: expected click rate
- Dataset partitioned into 5 big enough subgraphs
- Query set
  - Sampled from the same two-week period
  - Filter out the ones not present in subgraphs
  - 120 such queries
Experiments: Metrics

- Manual evaluation:
  - Manually assigned scored between 1-4 to every (query, rewrite) pair, by Yahoo! Team
  - Scores 1-2: relevant
  - Scores 3-4: irrelevant

- Precision:
  \[
  \text{precision}(q, m) = \frac{\text{relevant rewrites of } q \text{ that } m \text{ provides}}{\text{number of rewrites of } q \text{ that } m \text{ provides}}
  \]

- Recall:
  \[
  \text{recall}(q, m) = \frac{\text{relevant rewrites of } q \text{ that } m \text{ provides}}{\text{number of relevant rewrites of } q}
  \]
Experiments: Metrics

- **Query Coverage:**
  - Absolute number of queries for which there is at least one rewrite

- **Rewriting Depth:**
  - Number of rewrites for a given query
Experiments: Metrics

- Desirability:
  - Desirability function:
    \[
    des(q_1, q_2) = \sum_{i \in E(q_1) \cap E(q_2)} \frac{1}{|E(q_2)|} \cdot w(q_2, i)
    \]
  - If \( des(q_1, q_2) > des(q_1, q_3) \) then,
    \[ sim(q_1, q_2) > sim(q_1, q_3) \]
Results: baselines
Note: Pearson fares the best among baselines
Results: query coverage

Query Coverage for Pearson and variations of Simrank

- Weighted Simrank: 98%
- Evidence-based Simrank: 98%
- Simrank: 98%
- Pearson: 41%
Results: precision-recall

Precision-recall graphs

- ▲ Simrank
- ▼ Pearson
- O evidence-based Simrank
- * weighted Simrank

Precision

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Recall
Note: weighted simrank fares the best
Results: rewriting depth

Comparing the rewriting depth of Pearson and variations of Simrank

- Weighted Simrank
- Evidence-based Simrank
- Simrank
- Pearson

Percentage (%) of sample queries

# of rewrites:
- 5
- 4-5
- 3-5
- 2-6
- 1-6
Results: rewriting depth

Note: weighted simrank is among the best
Conclusion

- Two Simrank extensions
  - “evidence” supporting similarity
  - Weights of edges
- Weighted Simrank is overall the best
- Issues not addressed
  - Spam clicks
  - Semantic text-based similarities
  - Updating similarity scores with changes in click graph
Multi-partite?

What about more than two partitions?
THANKS