Contextual Advertising: Semantic Approach

Ekaterina Biehl
Overview:

Motivation: bit of history on Web monetization

Contextual advertising
  - Organisation
    - Types

Semantic Approach
  - Classification
    - Matching
      - Searching
    - Evaluation

Conclusion

WEB Advertising

- Banner ads
- Pop-up ads
- => software to eliminate from PCs
- Sponsored search-ads driven by originating query
- Contextual advertising (context match)
**Contextual Ads:**

**Definition**

**Context Match** refers to the placement of commercial textual advertisements within the content of a generic web page.

A **contextual ad** is the advertisement that dynamically appears on a Web site.

---

**Ads of sport-related companies:**
- sport equipment
- ticket sellers
Advertising: Organisation

Publisher -> AD network -> Advertiser

Users
TYPES of Contextual Ads

- **Search-based**: Google’s AdSense, Yahoo! Publisher Network

- **Channel-based**: Kanoodle, Valueclick

- **Behaviorally-based**: Tacoda, Blue Lithium

- **In-line Advertising**: Vibrant Media
Contextual Ads: Searching Formula

\[ \arg \max_i P(\text{click}|p, a_i) \]

\( p = \) given page; \( a = \) given ad
Syntactic vs. Semantic Approach

- Syntactic approach: estimates the ad relevance based on co-occurrence of the same words or phrases within an ad and a page

  the Chevy Tahoe Truck => Lake Tahoe vacations

- Semantic approach: combines a semantic phrase (classification of ads and pages into a taxonomy of topics) with traditional keyword matching

  the Chevy Tahoe Truck => automobile domain => Car/Truck ads
Taxonomy

6000 nodes

Each node: collection of around 100 exemplary bid phrases that correspond to the node concept

Idea: find page-ad pairs being topically close: classify pages and ads into the same taxonomy
Classification: Training data

- page training set: generate the top 10 results of the Web search index for each class in the taxonomy

- ad training set: select ads with a bid-phrase assigned to the class

- Use SVM and a log-regression classifiers

- => not good performance
Classification Method:

- Rocchio's nearest-neighbor classifier:
- Each taxonomy node: a single meta-document (concatenation of all the example queries), represented as
- a centroid for the class (sum of the $tf-idf$ values of each term)
- The classification is based on the cosine of the angle between the document and the centroid
Semantic-syntactic Matching

Convex combination of the keyword (syntactic) and classification (semantic) score:

\[
Score(p_i, a_i) = \alpha \cdot TaxScore(Tax(p_i), Tax(a_i)) + (1 - \alpha) \cdot KeywordScore(p_i, a_i)
\]

\(\alpha\) determines the relative weight of the taxonomy score and the keyword score
KeywordScore

- Pages and ads: vectors in n-dimensional space (one dimension for each term)
- The magnitude of each dimension: $tf \times idf$ score
- KeywordScore: the cosine of the angle between the page and the ad vectors
TaxonomyScore

Function:
- Topical match between a page and an ad
- Generalization within a taxonomy
- Efficient search of the ad space

match stronger ads and pages from the same node and weaker as the distance gets larger

Challenge: winter sport-> skiing, snowboarding hobby->sailing, knitting
Generalization

$$idist(c, p) = \frac{n_c}{n_p}$$

- $n_c$: Number of documents classified into the child node
- $n_p$: Number of documents classified into the child node
Searching: Inverted Index

- The ads are parsed into terms
- Each term has a weight based on a section where it appears
Searching: Inverted Index

- Challenge: how to preserve class information in the index
- Simple solution: unique meta-term for a class
  => loss of the generalization
- Instead: annotate each ad with one meta-term for each ancestor of the assigned class, weights of the meta-terms: the value of $idist()$ function
WAND is a document-at-a-time algorithm based on a two level approach:

- at the first level, it iterates in parallel over query term postings and identifies candidate documents using an approximate evaluation taking into account only partial information on term occurrences and no query independent factors; at the second level, promising candidates are fully evaluated and their exact scores are computed.
Evaluation

Percision vs. Recall of syntactic match vs. syntactic-semantic match
Impact of Alpha on precision for different levels of recall

**Precision Vs Alpha for Different Levels of Recall**
(Data Set 2)

- 80% Recall
- 60% Recall
- 40% Recall
- 20% Recall
Conclusion

Contextual Advertising: placement of commercial textual advertisements within the content of a generic web page

Approaches:
- Purely syntactic: keyword based
- Classification into a taxonomy
- Combination of keyword scores and semantic phrases (taxonomy scores)
Thank you!
Rocchio's Classifier

- Uses centroid vectors to represent a category
- Centroid vector is the average vector of all document vectors of a category
- Centroid vectors are calculated in the training phase
- To classify a new document, just calculate its distance to the centroid vector of each category
- Use cosine similarity as distance measure
- Advantages: fast training phase, fast classification
- Disadvantage: precision drops with increasing number of categories