CS834 - Introduction to Information Retrieval
Presentation #4

Hussam Aldeen Hallak

The Two Papers:

- A semantic approach to contextual advertising

  SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval
  Pages 559-566

- How much can behavioral targeting help online advertising?

  WWW '09 Proceedings of the 18th international conference on World wide web

  Pages 261-270

The First Paper:

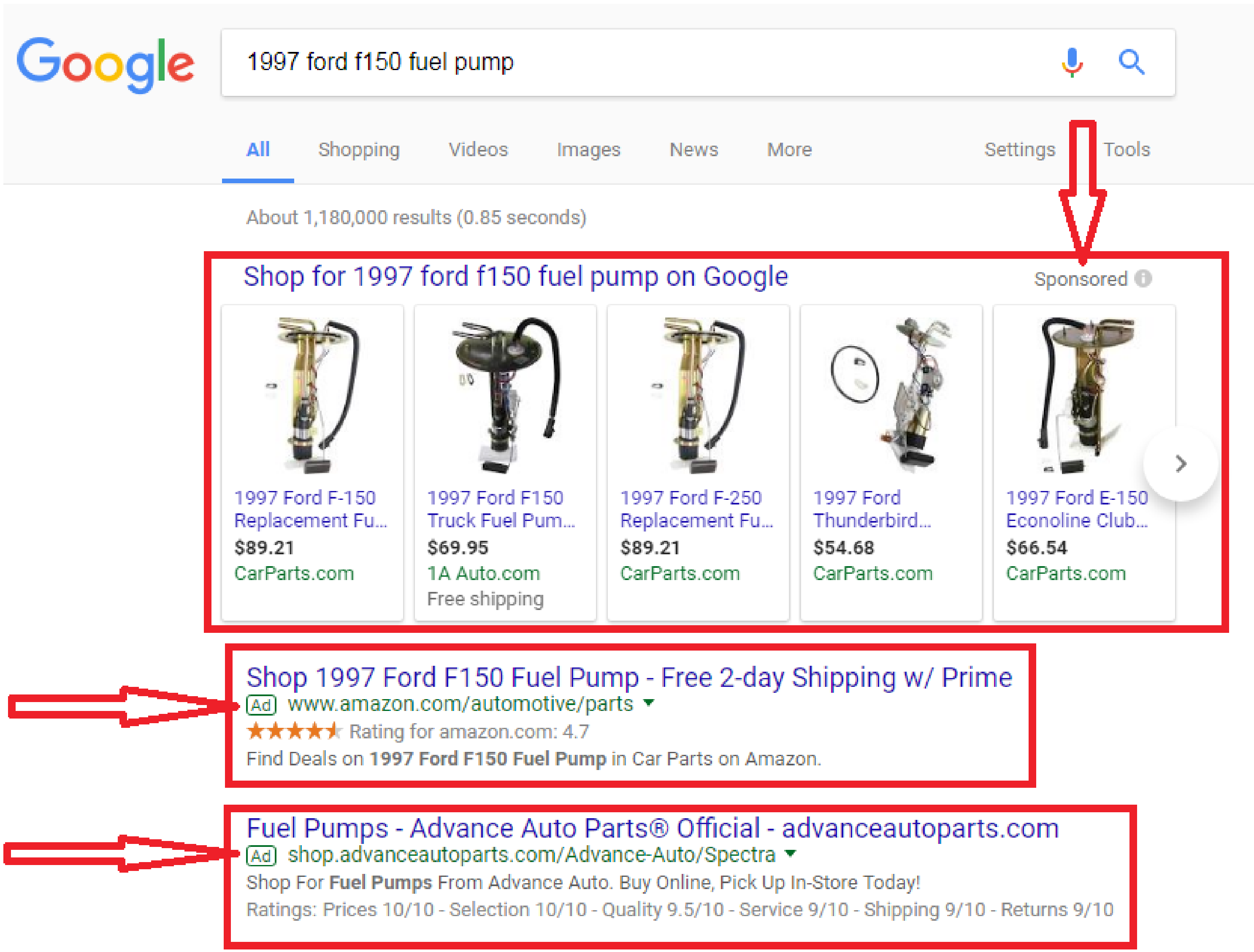
- A semantic approach to contextual advertising

  SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval
  Pages 559-566

http://clair.si.umich.edu/~radev/767w10/papers/Week12/ca/semantic.pdf

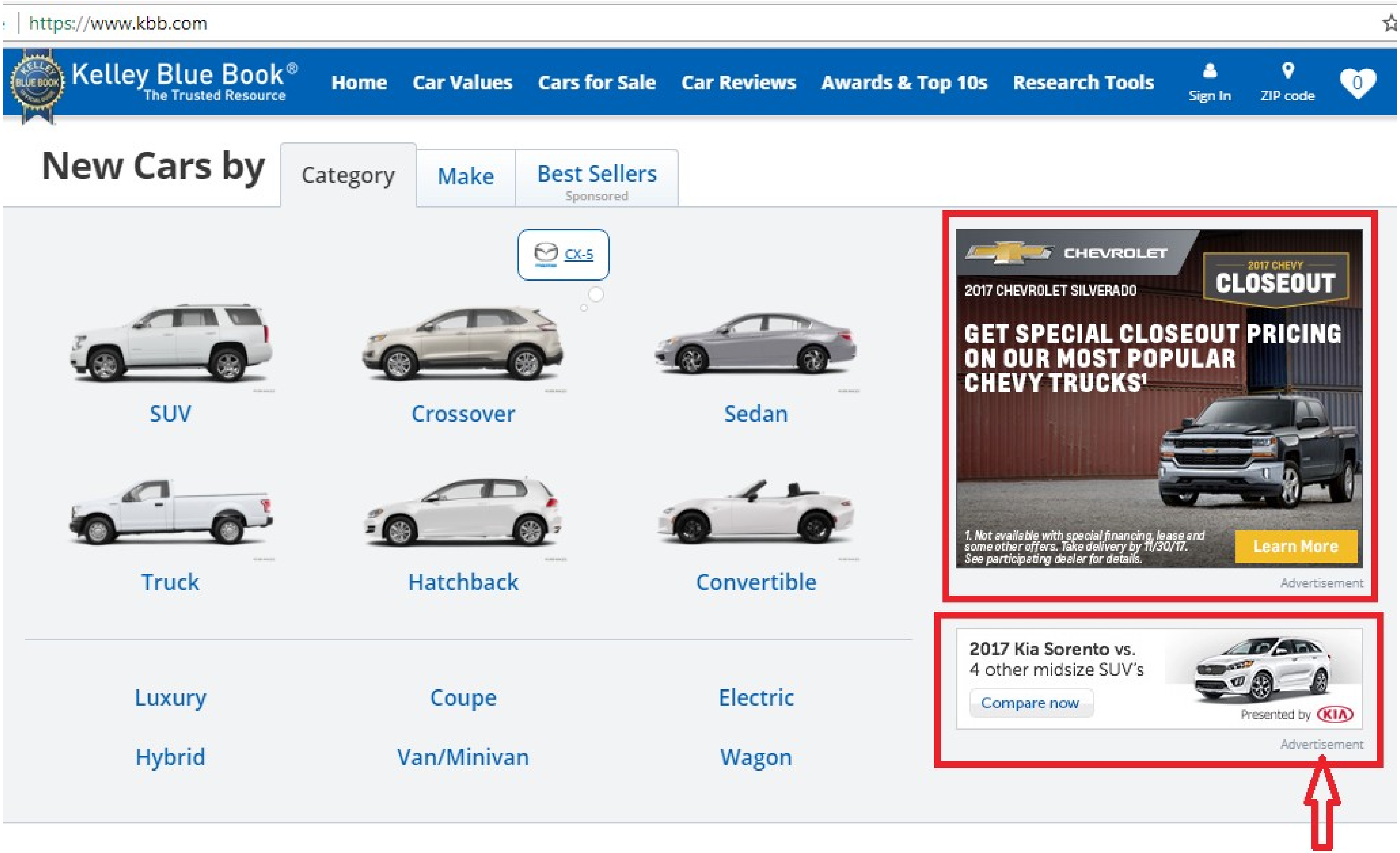
1st Type of Advertising Sponsored Search (SS)

- Placing ads on the result pages from a search engine



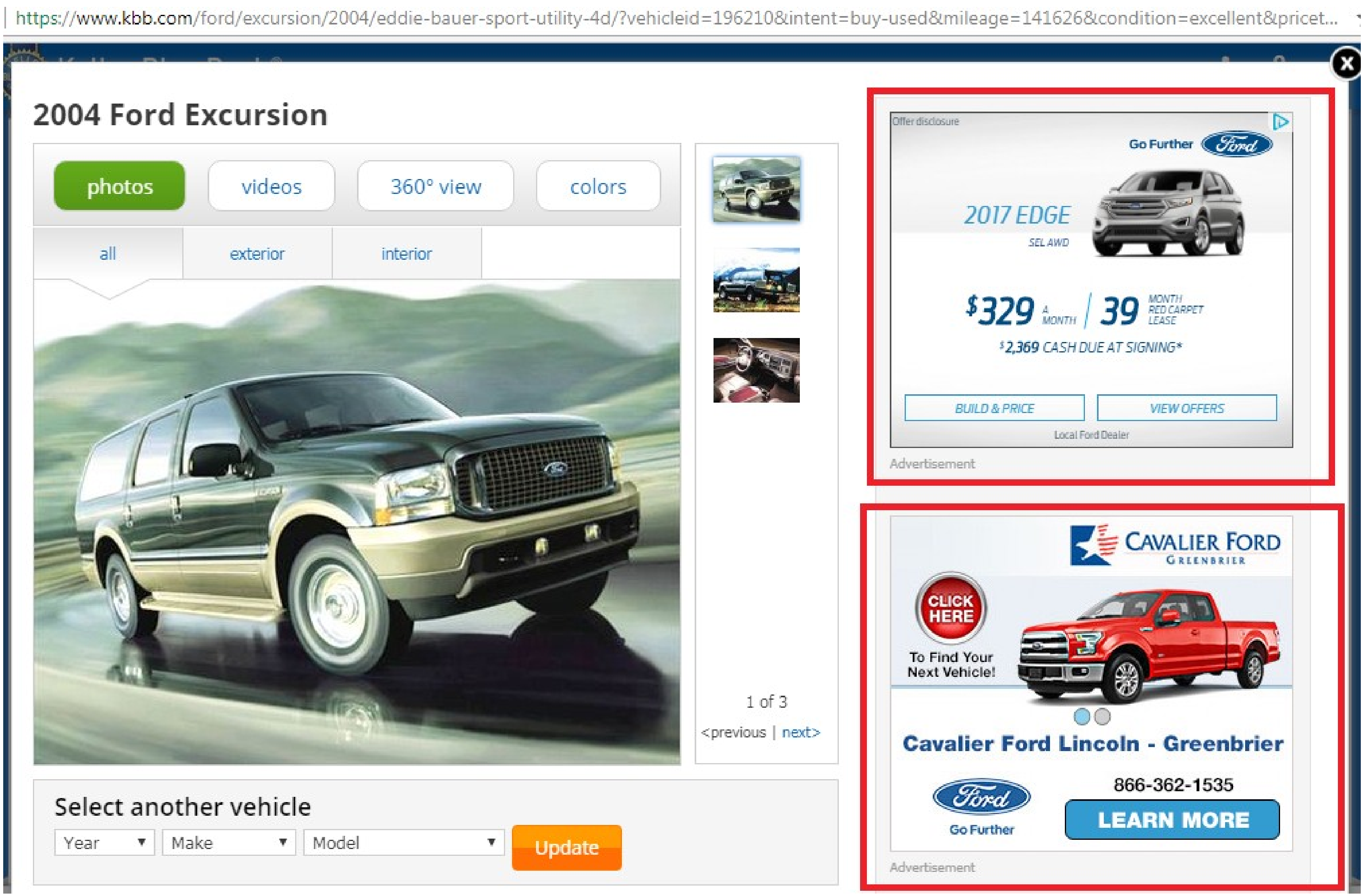
2nd Type of Advertising Context Match (CM)

- Commercial ads within the content of a web page.



Why ads and page content should be related?

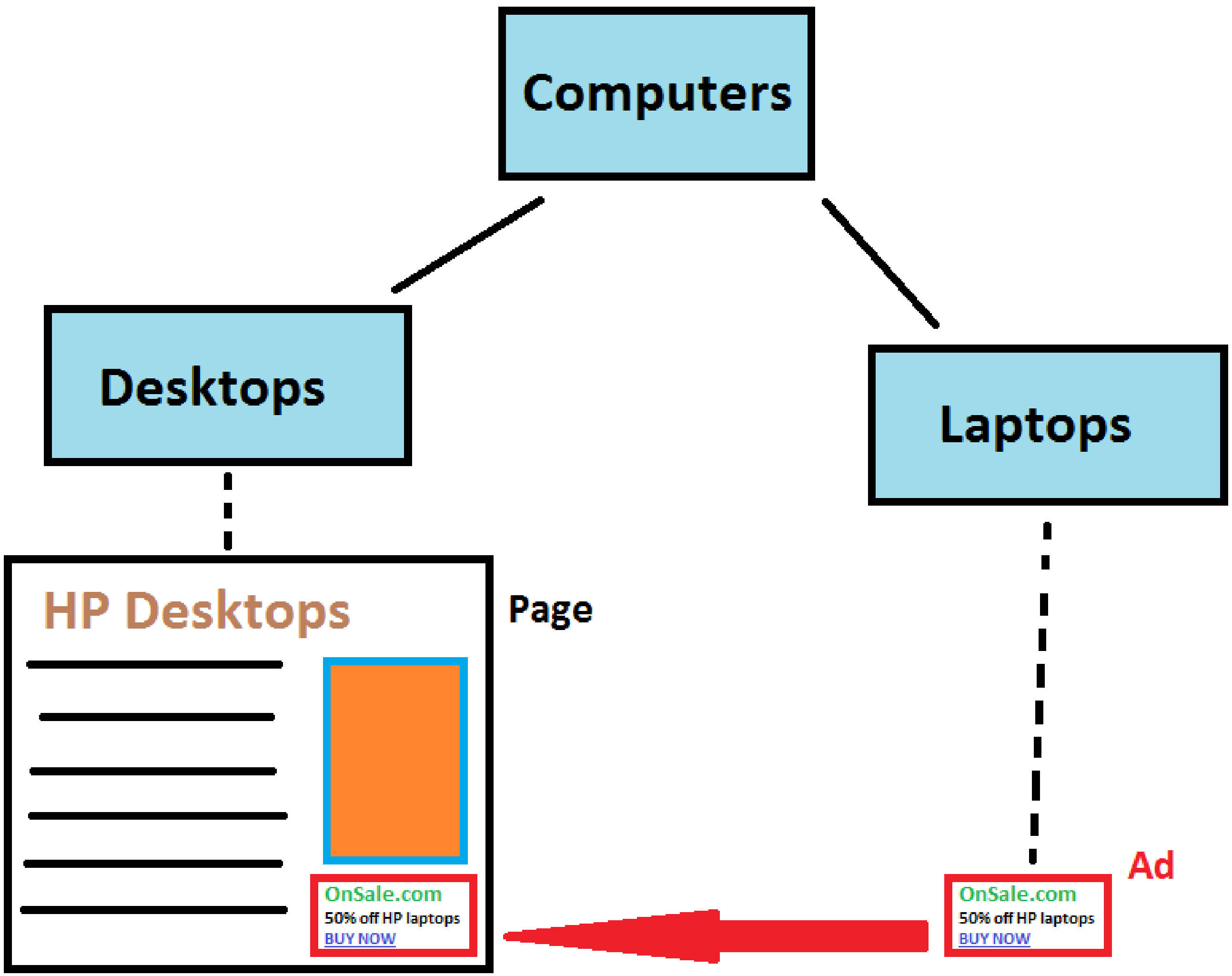
Improve user experience → increase clicks → increase revenue



Matching ads with pages

Syntactic Approach:
Match words found in the page with words in ads.
Problems:
Leads to irrelevant ads
A page about the golfer "John Maytag" might trigger an ad for "Maytag dishwashers"
Solution:
Combining Semantic (topical) and syntactic matching
The semantic phase:
Classify the page and the ads into a taxonomy of topics
Use the proximity of the ad and page classes as a factor in the ad ranking formula

Advantages of Using a hierarchical taxonomy



In some sense, the taxonomy classes are used to select the set of applicable ads and the keywords are used to narrow down the search.

Taxonomy Choice

Built by a large web search engine in the US
100 queries for each node
Contains 6000 nodes
Used for classifying both pages and ads

Categories per level

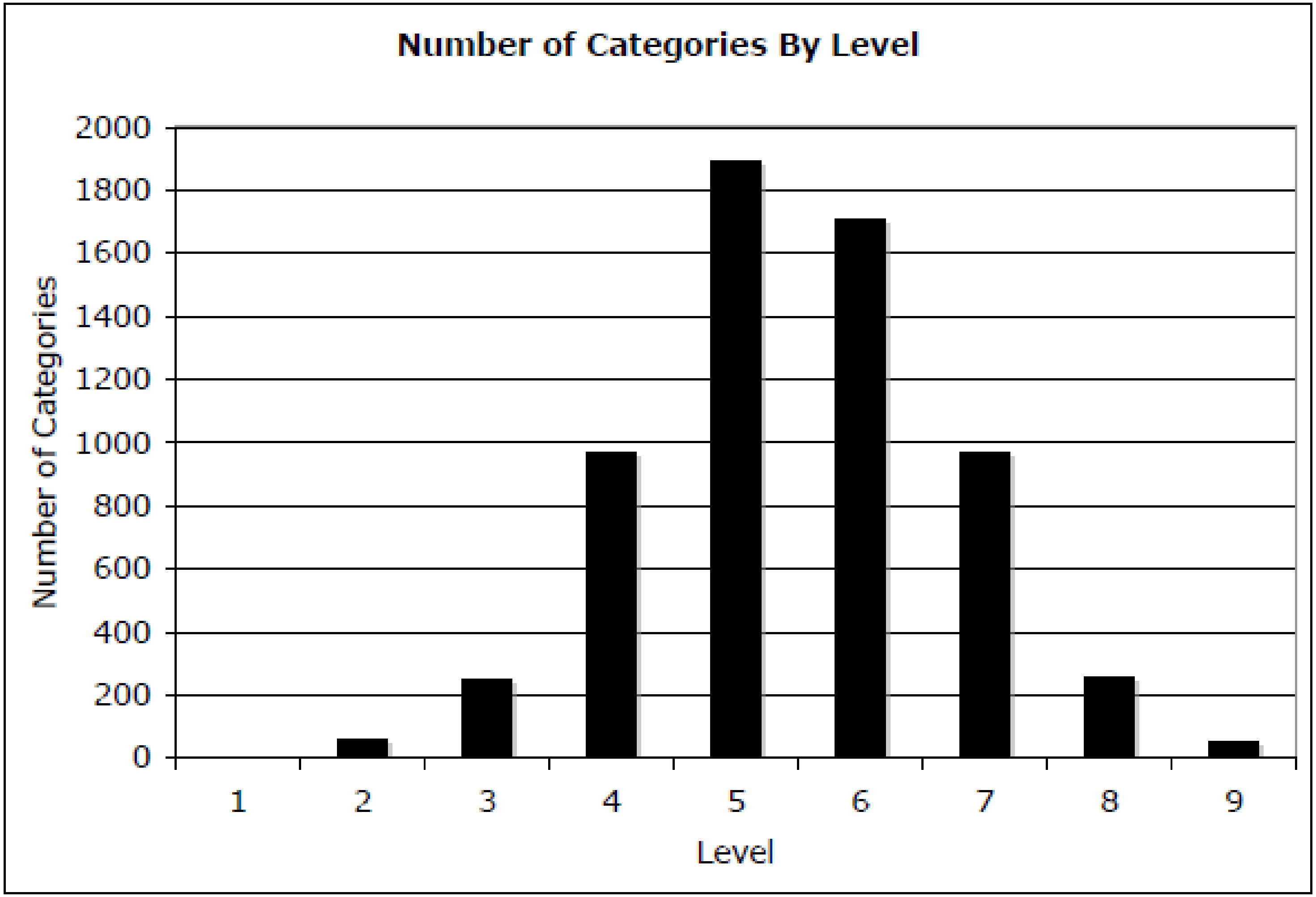


Figure 1: Taxonomy statistics: categories per level

Broder et al.(2007)

Number of children per node

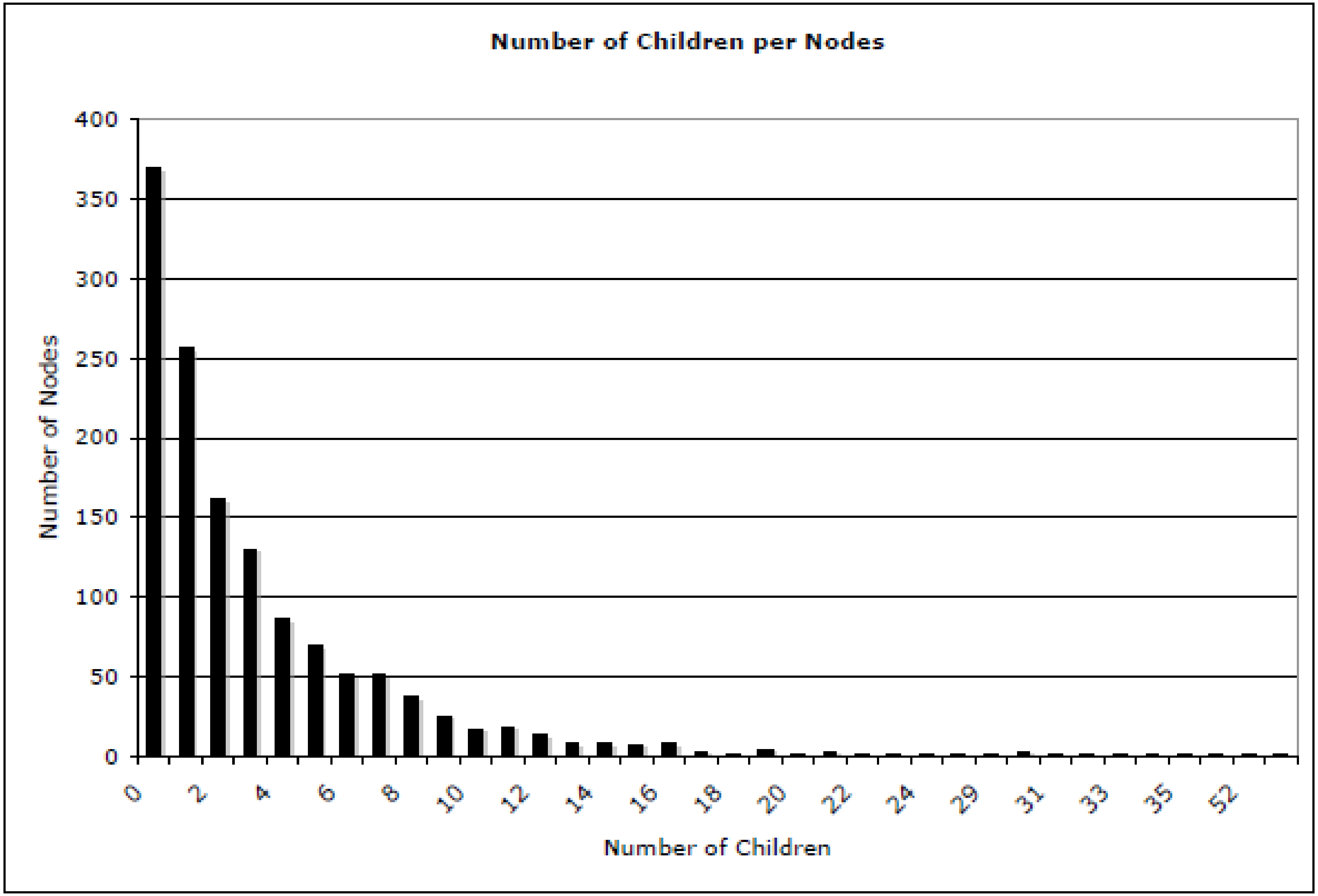


Figure 1: Taxonomy statistics: fanout for non-leaf nodes

Broder et al.(2007)

Queries per node

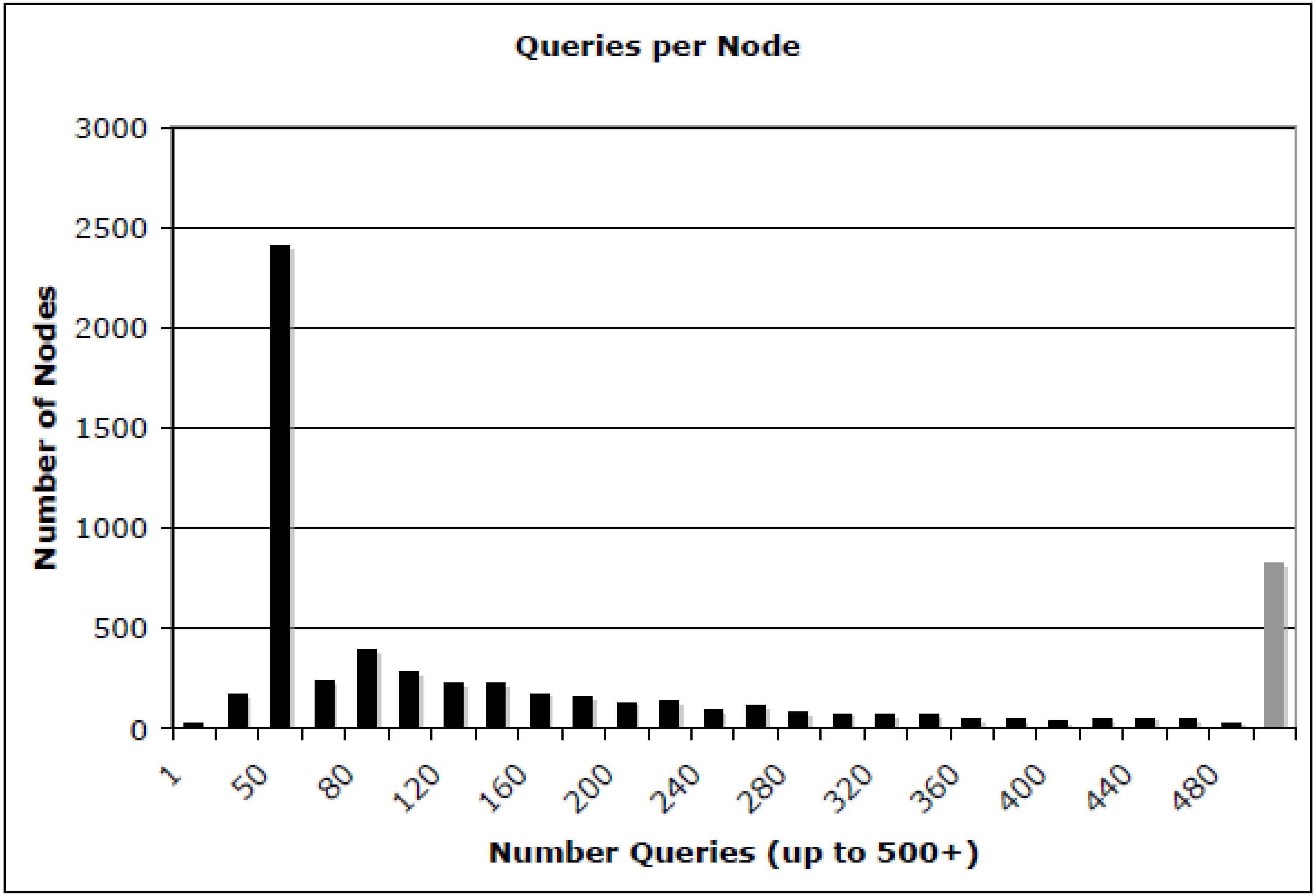


Figure 1: Taxonomy statistics: queries per node

Broder et al.(2007)

Training data

Pages: Select top 10 results of Web search index for each class in the taxonomy

Ads: Select ads with a bid-phrase assigned to the class

Classifiers:

SVM and log-regression classifiers were slow!

Rocchio's (nearest-neighbor) gave best performance!

Each taxonomy node: a single meta-document (concatenation of all the example queries), represented as a centroid for the class.

A centroid is the sum of the tf-idf values of each term.

$$\vec{c_j} = \frac{1}{|C_j|} \sum_{\vec{q} \in C_j} \frac{\vec{q}}{||\vec{q}||}$$

where $\vec{c}_j$ is the centroid for class C_j
q iterates over the queries in a particular class.

Classification is based on the cosine of the angle between the document and the centroid.

Semantic-Syntactic Matching

Process the content of the page

Extract features

Search the ad space to find the best matching ads.

Relevance score

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

$$Score(p_i, a_i) = \alpha \cdot TaxScore(Tax(p_i), Tax(a_i))$$

$$+(1-\alpha) \cdot KeywordScore(p_i, a_i)$$

α determines the relative weight of the taxonomy score and the keyword score.

α = TaxScore / KeywordScore

KeywordScore (syntactic relevance score)

Uses Vector Space Model

Pages and ads are vectors in n-dimensional Space (one dimension for each distinct term)

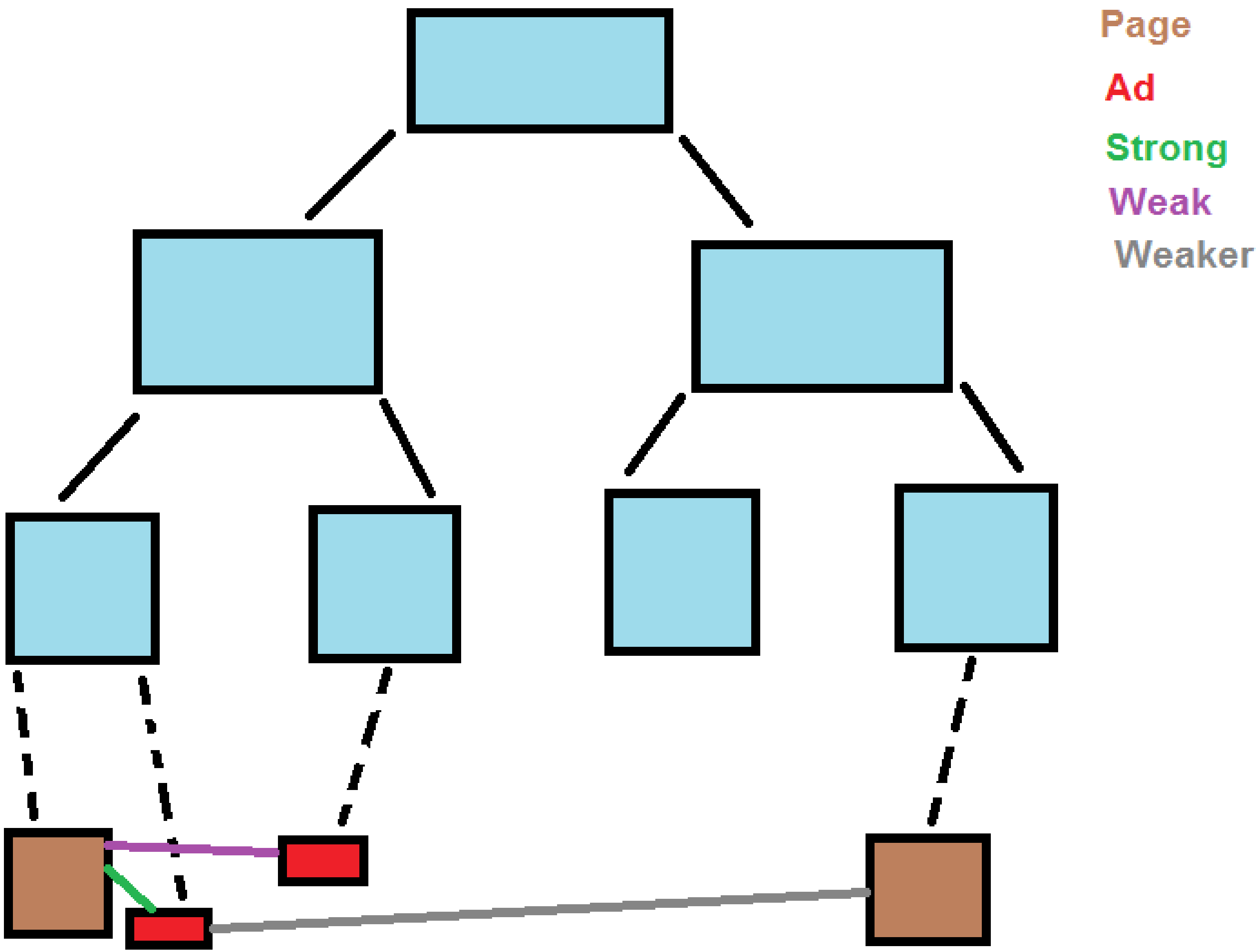
KeywordScore is the cosine of the angle between the page and the ad vectors

TaxonomyScore (semantic relevance score)

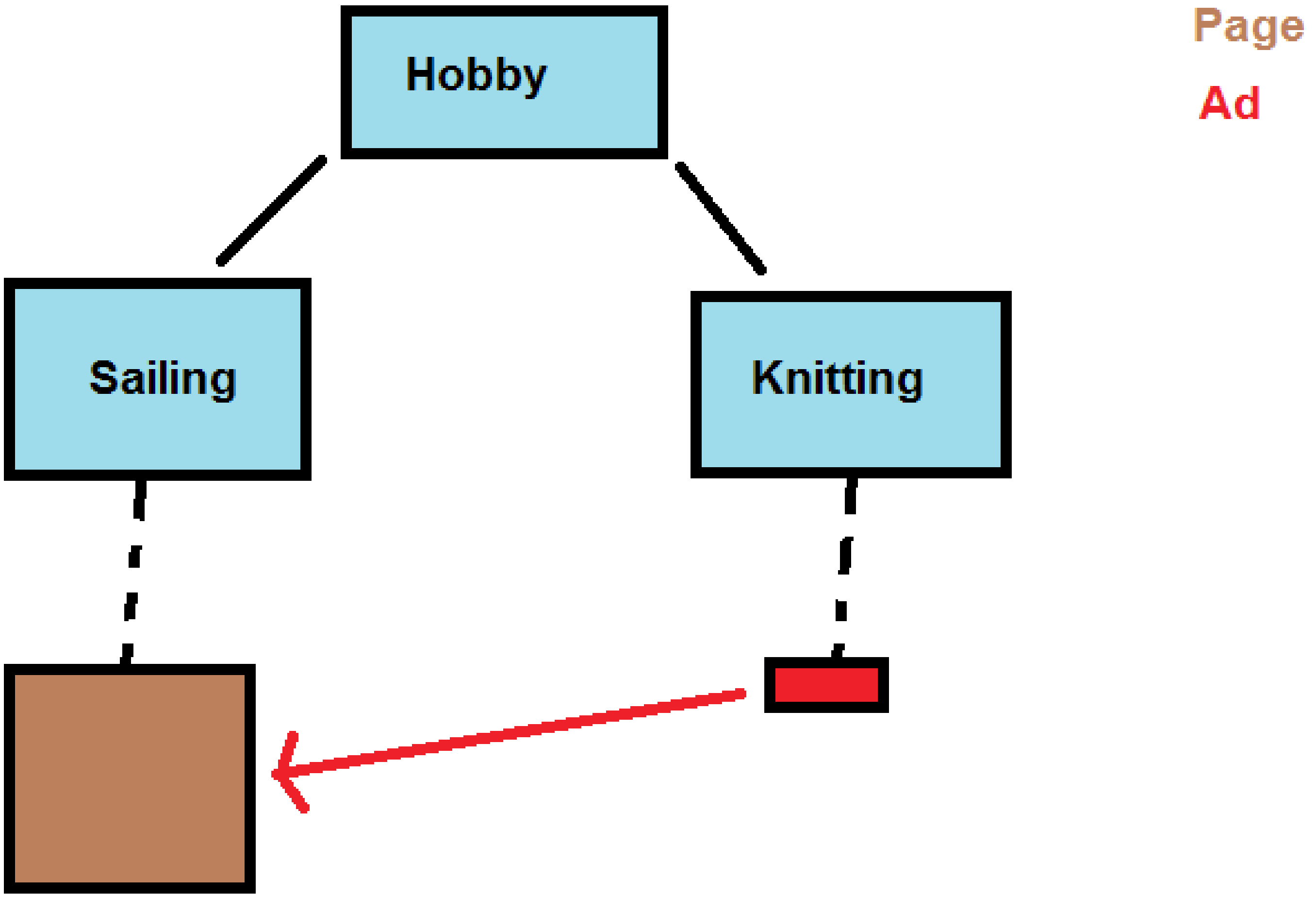
Purpose:
Match ads and pages based on the topic
Generalization within a taxonomy
Efficient search of the ad space (user is waiting)

Ideally:
The match is stronger when both the ad and the page are classified into the same node and weaker when the distance between the nodes in the taxonomy gets larger.

Ideal page-ad topical match



Generalization Challenge



Generalization Advantage

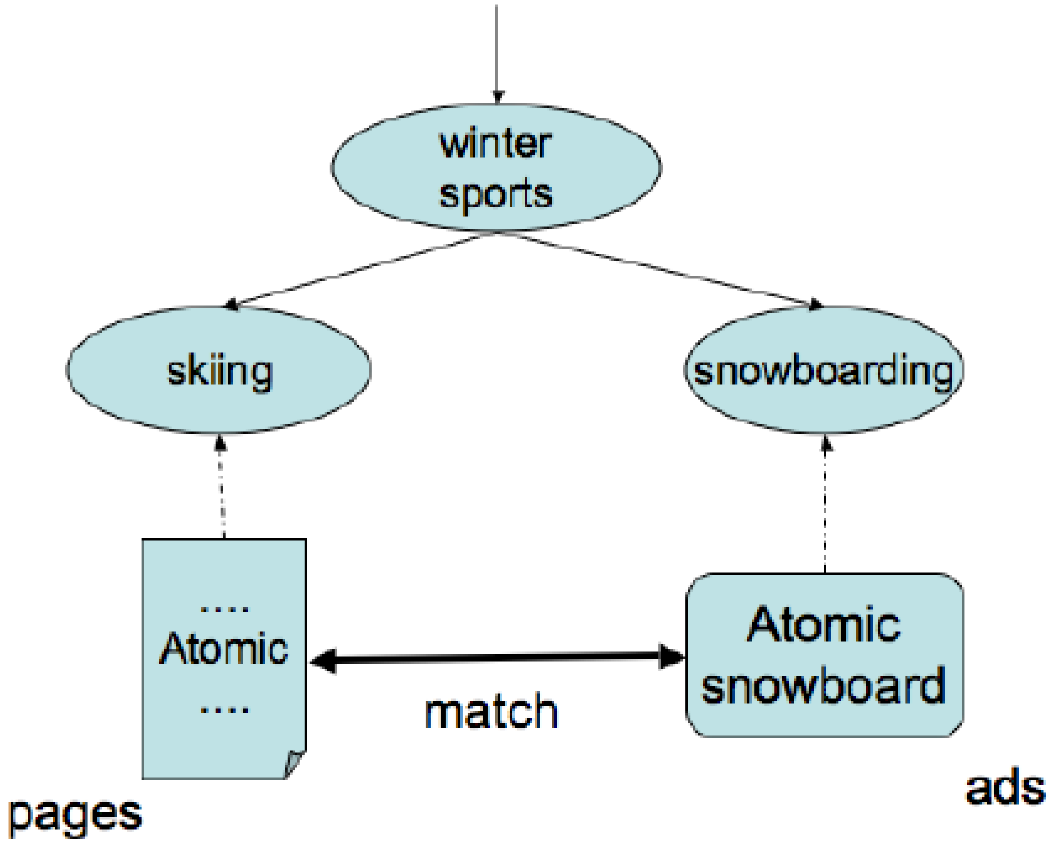


Figure 2: Two generalization paths

Broder et al.(2007)

Generalization cost

Density: The probability of an ad belonging to the parent topic being suitable for the child topic.

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

c is the child node.
p is the parent node.
n_c is the number of document classified into the subtree rooted at c.
n_p is the number of document classified into the subtree rooted at p.

0 ≤ idist ≤ 1
idist = 1 (the page and ad belong to the same class/node)
idist = 0 (the least common ancestor of page and ad is the root)

Searching the ad space

Using inverted index

The ads are parsed into terms

Each term has a weight based on a section where it appears

Challenge: Preserving class information in the index

Solution: annotate ads with a unique meta-term for each class

Cons: Generalization is lost!

Instead: Also annotate each ad with one meta-term for each ancestor of the assigned class; utilize weights!

Weights of the meta-terms: the value of idist() function

Example:

Atomic, skii, snow

{skiing, winter sport, sport}

index {
atomic: ad1, ad2
snow: ad2
skiing: ad1
snowboarding: ad2
wintersport: ad1, ad2
}

Atomic, skii, snow

{skiing, winter sport, sport}

index {
atomic: ad1, ad2
snow: ad2
~~skiing: ad1~~
snowboarding: ad2
wintersport: ad1, ad2
}

Data and Methodology

- Data: 105 pages randomly selected from 20 million pages with contextual advertising
- Tens of millions of ads from advertising network in the US
- Human judges for each page-ad pair on a 1 to 3 scale:

  1. Relevant

  Page: The National Football League

  Ad: Tickets for NFL games

  2. Somewhat Relevant

  Page: The National Football League

  Ad: NFL branded products

  3. Irrelevant:

  Page: The National Football League

  Ad: NFL player “John Maytag” triggers “Maytag” dishwasher ads on NFL page.

Results

pages	105
words per page	868
judgments	2946
judg. inter-editor agreement	84%
unique ads	2680
unique ads per page	25.5
page classification precision	70%
ad classification precision	86%

Table 1: Dataset statistics

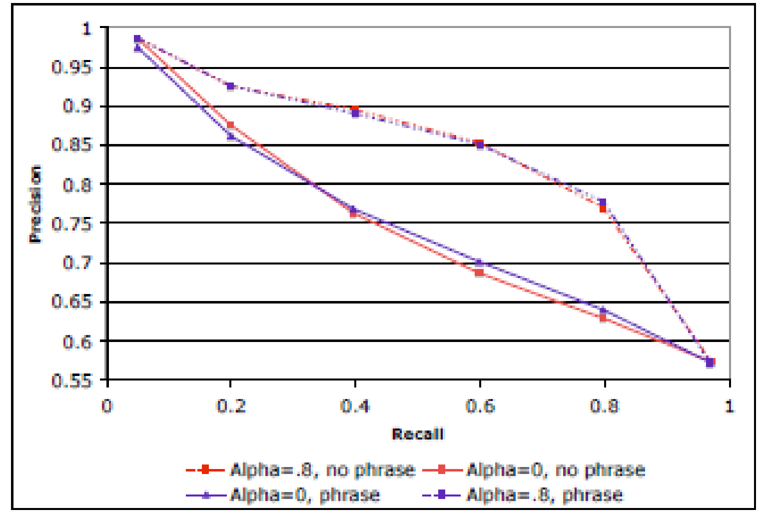


Figure 3: Data Set 2: Precision vs. Recall of syntactic match ($\alpha = 0$) vs. syntactic-semantic match ($\alpha = 0.8$)

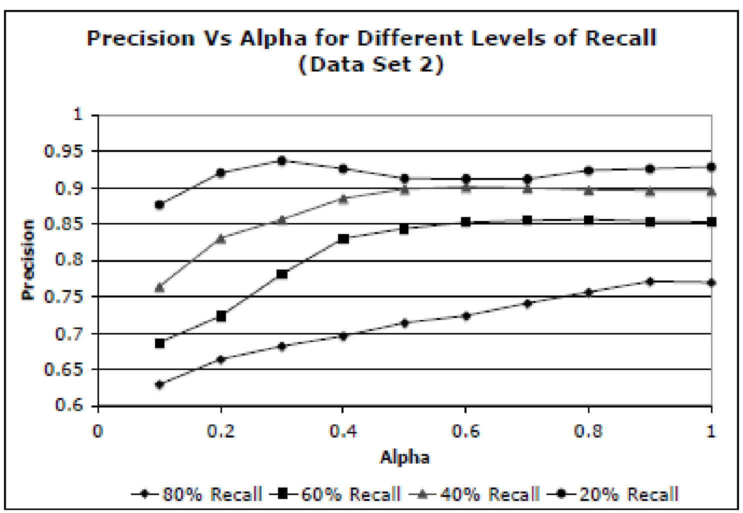


Figure 4: Impact of α on precision for different levels of recall

In most cases, precision grows or is flat when Alpha is increased, except at the low level of recall where due to small number of data points there is a bit of jitter in the results.

What to take away?

Syntactic (keyword) matching between pages and ads leads to irrelevant ads.

Semantic (topical) matching relies on the matching between pages and ads in topic

Semantic matching complemented with syntactic matching leads to better results.

The Second Paper:

- How much can behavioral targeting help online advertising?

WWW '09 Proceedings of the 18th international conference on World wide web

Pages 261-270

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.215.1473&rep=rep1&type=pdf

What is Behavioral Targeting BT Ads?

- The delivery of ads to targeted users based on their web search and browsing history.
- How much can BT help online advertising?
- Strategies to represent the users’ behavior:
  1. Web browsing history
  2. Search queries
- Which BT strategy is better for user segmentation?
- The performance of online advertising is measured by ads Click-Through Rate (CTR)
- Calculate & compare ads click entropy, precision, recall, and F-measure for different BT strategies.
- Short window (1 day) vs long window (7 days)

Modeling User Browsing History

- Users are represented by a matrix U_{gxl}
  g: number of users
  l: number of urls
- TF.IDF:
  Users are the documents
  URLs are the terms
- Each entry in the matrix is given the value:

	Lnk_1	Lnk_2	...	Lnk_l
usr_1				
usr_2				
...				
usr_g				

$$u_{ij} = \log[count(clicks\ on\ URL_j\ by\ User_i) + 1] \times \log[\frac{l}{count(users\ who\ have\ clicked\ on\ URL_j)}]$$

Modeling User Search History

- Query history uses Bag-of-Words (BOW) model to populate TF.IDF matrix.
- Stop words are removed and Porter stemming is used
- Terms that only appear once are removed. (why?)
- Number of terms is reduced from 765k to 294k

Examined BT Strategies:

- Two BT strategies (pages visited, queries searched)
- Two window sizes (long term: 7 days, short term: 1 day)
- Lead to four possible BT strategies to assess:
  1. Long term behavior based on page views (LP)
  2. Long term behavior based on query terms (LQ)
  3. Short term behavior based on page views (SP)
  4. Short term behavior based on query terms (SQ)

Raw Data Clean-Up!

- Source: A commercial search engine (Bing?)
- log dataset records all users' search click behavior: 1. Web page clicking
  2. Ad clicking
- Period: 7 days' click-through log data ranging from June 1st to 7th 2008
- Users removed if they clicked on more than 100 ads in one day (robots)
- Ads removed if they have fewer than 30 clicks over seven days (cannot be used to draw a reliable statistical conclusions)
- Only English queries are considered!

Click-Through Log Format

Table 1. Format of click-through log used in our study.

UserID	UID030608473X	A user ID for each unique user.
QueryText	xbox	The detailed query text used by the user
QueryTime	08-06-03 21:15:47	The time when the query was issued
ClickTime	08-06-03 21:16:02	The time when the click occurred after the query was issued
ClickURL	http://www.xbox365.com	The URL which has been clicked by the user
IsAd	0	A Boolean value to show the clicked URL is an ad or not
NumberAd	3	The number of ads displayed in the search results
DisplayAd	http://video-games.half.ebay.com/ http://accessories.us.dell.com/ http://www.gamefly.com	The URL list of all the ads that displayed by the query. (To save space, we only reserve top domain of the ad URL in this example.)

Definitions:

- $A = \{a_1, a_2, \ldots, a_n\}$ is the set of ads
- $Q_i = \{q_{i1}, q_{i2}, \ldots, q_{imi}\}$ queries which have displayed or clicked a_i
- $U_i = \{U_{i1}, U_{i2}, \ldots, U_{imi}\}$ users who have displayed or clicked a_i
- $\delta(u_{ij})$ is used to show whether the user u_{ij} has clicked the ad a_i

$$\delta(u_{ij}) = \begin{cases} 1 & \text{if } u_{ij} \text{clicked } a_i \\ 0 & \text{otherwise} \end{cases}$$

$l_i = \sum_j \delta(u_{ij})$: the number of users clicked ad a_i

- K-means and and CLUTO (a clustering software package) cluster users into groups: $g_k(U_i)$ is all users in U_i

$$G(U_i) = \{g_1(U_i), g_2(U_i), \ldots, g_K(U_i)\}, i{=}1,2,\ldots n$$

Calculating Similarity Between Users

- Cosine similarity is used:

$$Sim(u_{ij}, u_{st}) = \frac{< u_{ij}, u_{st} >}{||u_{ij}||\,||u_{st}||}$$

- Within-ad similarity: Users who clicked the same ad

$$S_w(a_i) = \frac{2}{l_i(l_i - 1)} \sum_{\delta(u_{ij})=1} \sum_{\substack{\delta(u_{it})=1 \\ t \neq j}} Sim(u_{ij}, u_{it})$$

- Between-ad similarity: Users who clicked different ads

$$S_b(a_i, a_s) = \frac{1}{l_i l_s} \sum_{\delta(u_{ij})=1} \sum_{\delta(u_{st})=1} Sim(u_{ij}, u_{st})$$

Similarity Ratio

- Ratio between within-ad and between-ad similarity

$$R(a_i, a_s) = \frac{S_w(a_i) + S_w(a_s)}{2S_b(a_i, a_s)}$$

- The larger the ratio, the more confident we are on the basic assumption of BT for a pair of ads a_i and a_s

Ad Click-Through Rate (CTR)

- The CTR of ad a_i is defined as the number of users who clicked it over the number of users who either clicked it or only displayed it.

$$CTR(a_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} \delta(u_{ij})$$

F-measure

- CTR can be used to calculate precision and recall
- Positive instance: Users displayed and clicked a_i
- Negative instance: Users displayed a_i but didn't click it
- Precision: CTR of segment $Pre(a_i|g_k) = CTR(a_i|g_k)$
- Recall: clicks of segment/total clicks $Rec(a_i|g_k) = \frac{\sum_{u_{ij} \in g_k(U_i)} \delta(u_{ij})}{\sum_{j=1}^{m_i} \delta(u_{ij})}$
- F-measure:

$$F(a_i|g_k) = \frac{2Pre(a_i|g_k)Rec(a_i|g_k)}{Pre(a_i|g_k) + Rec(a_i|g_k)}$$

- The larger the F measure is, the better the achieved performance is by user segmentation for BT

Ads-Click Entropy

- For ad a_i, the probability of users in segment g_k, who will click this ad, is estimated by:

$$P(g_k|a_i) = \frac{1}{m_i} \sum_{u_{ij} \in g_k(U_i)} \delta(u_{ij})$$

- Ads-Click Entropy (mathematically):

$$Enp(a_i) = -\sum_{k=1}^{K} P(g_k|a_i) log P(g_k|a_i)$$

- Smaller Entropy => Better user segmentation

Within- and between- ads user similarity

Table 2. Within- and between- ads user similarity.

	S_w	S_b	R
LP	0.1417	0.0252	28.9217
LQ	0.2239	0.0196	44.2908
SP	0.1532	0.0281	24.5086
SQ	0.2594	0.0161	$\boxed{91.1890}$

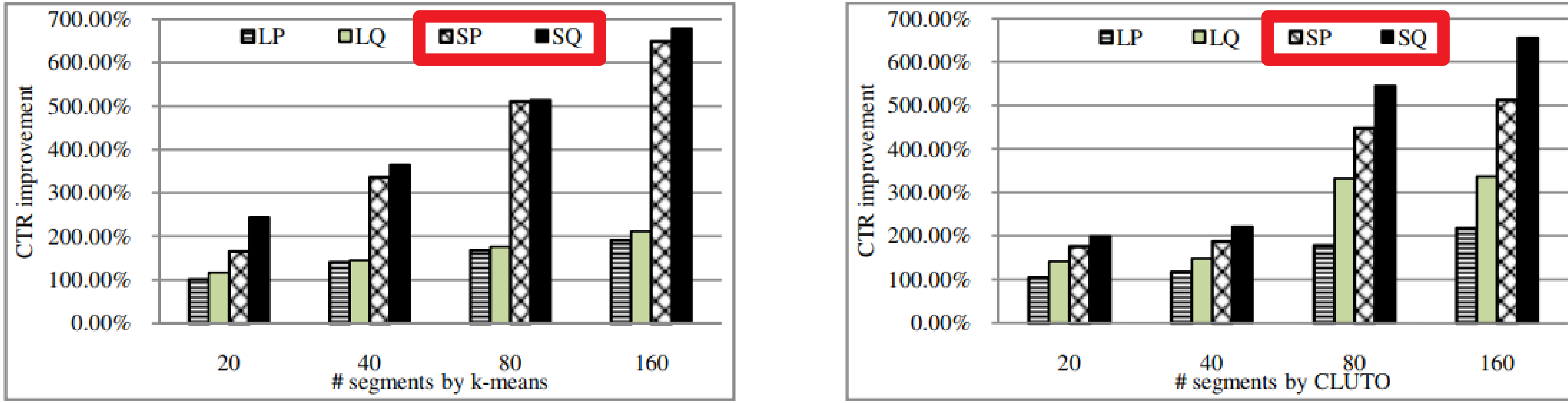
Scores are average across all ads or query terms

- Users who clicked the same ad are up to 91 times more similar

- For all ad pairs, 99.37% had higher within-ad user similarity than between-ad similarity.

- Search queries are more effective than pages clicked in BT

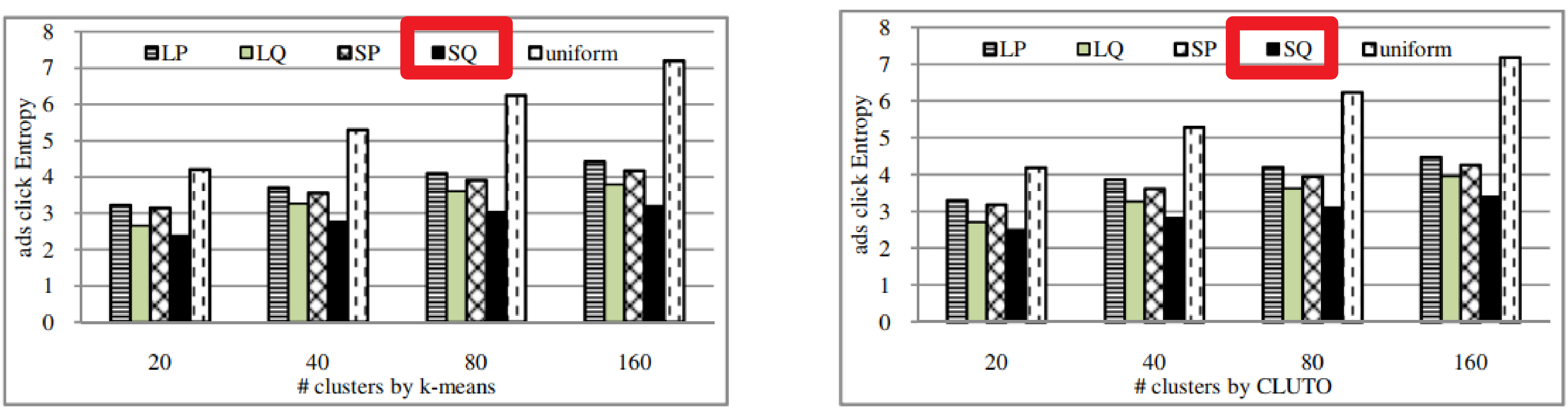
User segmentation improves CTR by 670%:



(a) User clustering by k-means (b) User clustering by CLUTO

Figure 1. CTR improvements by user segmentation for BT.

- CTR was improved by up to 670% off of the non segmented CTR



(a) Cluster by k-means (b) Cluster by CLUTO

Figure 2. Ads click Entropy of user segmentation for BT.

Overall Performance by different methods:

Table 5. F measure of different BT strategies

		LP	LQ	SP	SQ
K-means (20 segments)	*Pre*	8.67%	8.60%	13.35%	**17.08%**
	Rec	10.20%	22.34%	7.63%	**25.58%**
	F	0.08	0.10	0.08	**0.16**
CLUTO (20 segments)	*Pre*	8.62%	8.56%	14.61%	**19.13%**
	Rec	10.01%	20.51%	7.86%	**21.43%**
	F	0.08	0.10	0.07	**0.15**
K-means (40 segments)	*Pre*	8.84%	9.23%	19.76%	**20.53%**
	Rec	9.48%	18.20%	4.83%	**20.75%**
	F	0.08	0.10	0.06	**0.16**
CLUTO (40 segments)	*Pre*	8.76%	9.14%	19.38%	**22.80%**
	Rec	8.44%	**17.88%**	4.52%	17.78%
	F	0.08	0.10	0.06	**0.14**
K-means (80 segments)	*Pre*	9.02%	9.63%	23.47%	**23.49%**
	Rec	8.93%	17.62%	4.06%	**19.35%**
	F	0.08	0.10	0.06	**0.16**
CLUTO (80 segments)	*Pre*	8.85%	9.51%	23.09%	**27.00%**
	Rec	7.82%	**16.65%**	4.00%	15.55%
	F	0.07	0.10	0.06	**0.15**
K-means (160 segments)	*Pre*	9.09%	9.93%	25.68%	**25.81%**
	Rec	8.54%	17.98%	3.92%	**19.78%**
	F	0.074	0.10	0.06	**0.17**
CLUTO (160 segments)	*Pre*	8.87%	9.84%	25.43%	**31.02%**
	Rec	7.24%	**15.58%**	3.78%	14.52%
	F	0.07	0.10	0.06	**0.15**

Precision vs Recall for 160 user segments

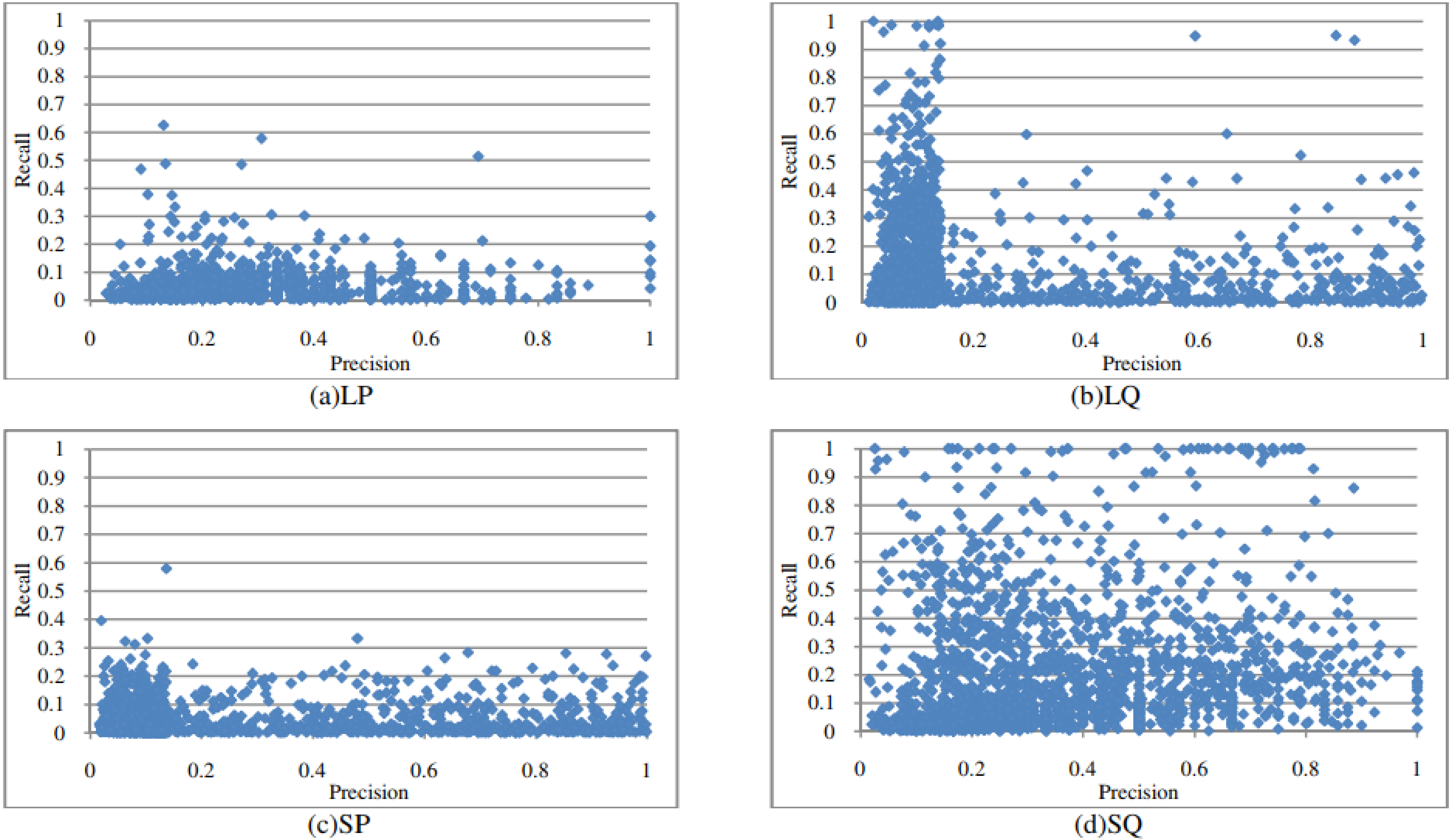


Figure 3. Scatter plot of Precision and Recall over all the ads (CLUTO-160 user segments).

Take away message:

- Behavioral targeting can help online advertisement
- Users who click the same ads have similar behavior and they are 91 times more similar than those who don't
- User segmentation based on search queries gives better results than pages visited
- It is better to use short window to segment users
- Increasing the number of segments provides better targeting.
- CTR can be improved by 670% using user segmentation

Referrences:

- A semantic approach to contextual advertising

  SIGIR '07 Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval Pages 559-566

  http://clair.si.umich.edu/~radev/767w10/papers/Week12/ca/semantic.pdf

- How much can behavioral targeting help online advertising?

  WWW '09 Proceedings of the 18th international conference on World wide web

  Pages 261-270

  http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.215.1473&rep=rep1&type=pdf

Why did the 2nd paper cite the 1st paper?

- Start here