Query-log mining for detecting spam queries

Carlos Castillo[1], Claudio Corsi[2], Debora Donato[1],
Paolo Feraggina[2], Aristides Gionis[1]

[1]Yahoo! Research Labs, Barcelona, Spain
[2]University of Pisa, Italy

motivation

- Query logs provide valuable information for queries and for documents
  - implicit tags
  - wisdom of crowds
- Human-constructed directories provide high quality classification labels for (a subset) of douments

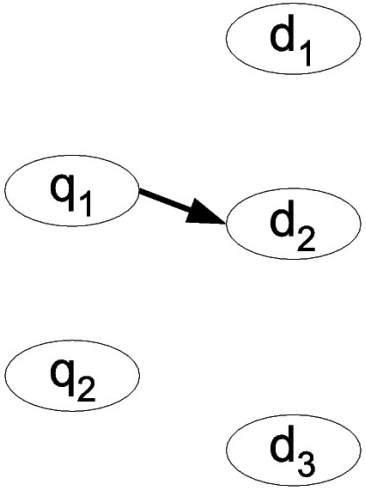
$\Rightarrow$ Identify spam by combining information contained in query logs and in web directories and usage mining

main idea

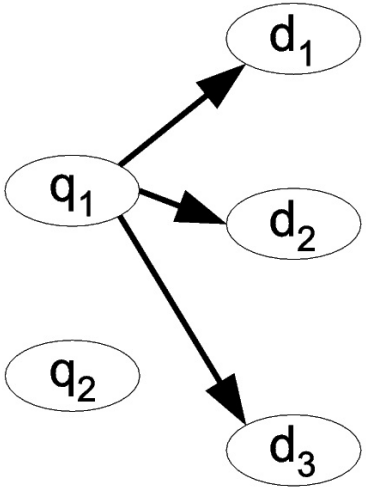
- *Query graphs:* bipartite graphs between queries and documents
- Extract features from query graphs
- "Semantic" features obtained by propagating web-directory topic labels on the query graph
- Use obtained features to improve accuracy of spam detection
- Characterize also queries as *spam-attracting*

click graph, view graph, and anticlick graph

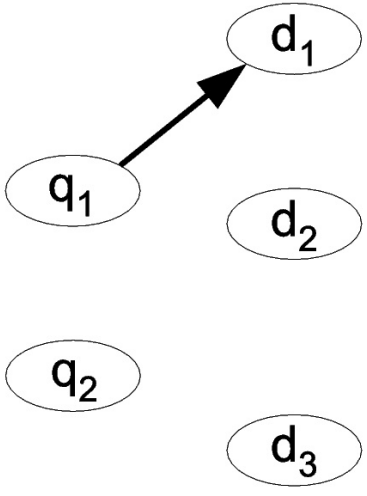
Example query log entry:
q_1: "*shoes*" => d_1: shoes1.com, d_2: shoes2.com [clicked], d_3: shoes3.com



Click-Graph



View-Graph



Anticlick-Graph

syntactic features

- degree of a node (query or document)
- for document d: $\texttt{topQ}_x(d)$ the set of queries adjacent to d and being among the fraction x of the most frequent queries in the query log
- for document d: $\texttt{topT}_y(d)$ the set of query terms which compose the queries adjacent to d in G and being among the fraction y of the most frequent terms in the query log

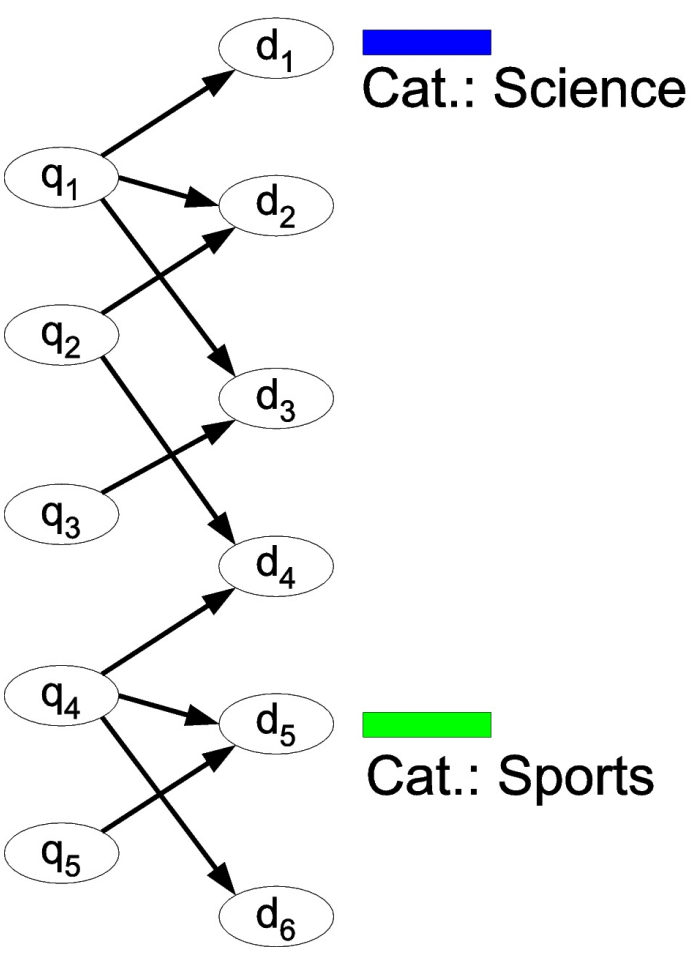
topics

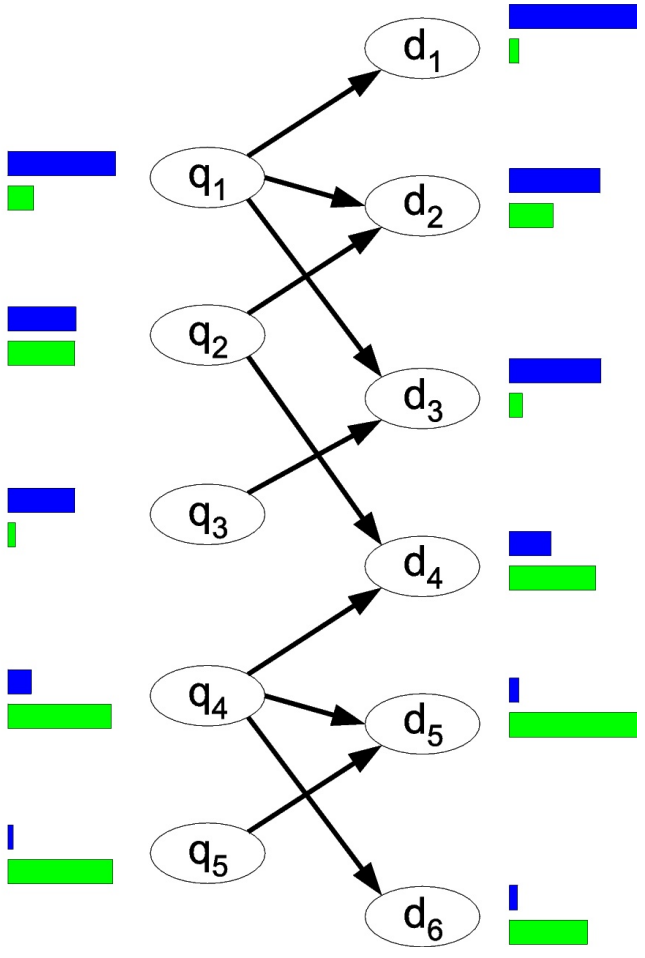
- intuition: multi-topic attractor has potential of being spam
- topic labels can be obtain from a web directory
- ...but not for all documents

topics

- intuition: multi-topic attractor has potential of being spam
- topic labels can be obtain from a web directory
- ...but not for all documents

propagation





Read result at each node as a distribution, and compute its entropy

propagation

- propagation by weighted average

$$\texttt{score}_v^{i+1}(c) \;\; += \;\; \alpha^{i-1} \sum_{(v',v)\in E} \texttt{score}_{v'}^{i}(c) \times f(v',v)$$

  and normalization

- propagation by random walk
  - inspired by topic-sensitive PageRank

- "Semantic features": *entropy* of the distribution of topic scores (documents and queries)

datasets

- query-log: sample of 1.6m queries from Yahoo! query log
- web dirctory: DMOZ, 4.2m documents
- labeled spam colection: the `WEBSPAM-UK2006` dataset

statistics on the query graphs

	Document-level			Host-level		
	C_d	A_d	V_d	C_h	A_h	V_h
Queries	1.59M	0.75M	2.78M	1.59M	0.75M	2.78M
Docs/hosts	2.75M	1.31M	23.47M	0.83M	0.40M	3.08M
Edges	3.69M	1.67M	40.71M	3.50M	1.53M	3.45M
$\mathcal{C}_D(0)$	0.05	0.08	0.03	0.28	0.35	0.15
$\mathcal{C}_Q(1)$	0.18	0.24	0.39	0.58	0.75	0.92
$\mathcal{C}_D(2)$	0.22	0.22	0.45	0.70	0.75	0.94
$\mathrm{CC}_{\max}$	0.32	0.19	0.92	0.80	0.83	0.98
$\lvert\mathrm{CC}\rvert$	0.21	0.23	0.007	0.08	0.06	0.006

finding web spam

Feature set	Features	TP	FP	F_1	AUC
Content (C)	98	75.8%	9.8%	0.692	0.912
Links (L)	139	84.2%	9.5%	0.739	0.939
Usage (U)	61	54.2%	7.4%	0.557	0.872
C ∪ L	237	83.9%	8.6%	0.756	0.952
C ∪ U	159	68.4%	6.6%	0.693	0.917
L ∪ U	200	78.5%	6.5%	0.757	0.951
C ∪ L ∪ U	298	78.9%	6.2%	0.765	0.951

finding spam-attracting queries

- define "spamicity of a query": fraction of spam results shown to the user
- Task 1: predict if query spamicity is "< 0.5" or "≥ 0.5"
  AUC: 0.798, true positive rate: 73.7%, false positives: 29.0%
- Task 1: predict if query spamicity is "$= 0.5$" or "≥ 0.5"
  AUC: 0.838, true positive rate: 74.0%, false positives: 22.1%

summary

- Use query-log mining and DMOZ class labels for spam detection
- Detect spam that has already “fooled” the search engine
- Propagation method can be useful in other tasks, too
- Future: extract better features and improve the results

summary

- Use query-log mining and DMOZ class labels for spam detection
- Detect spam that has already “fooled” the search engine
- Propagation method can be useful in other tasks, too
- Future: extract better features and improve the results

Thank you!