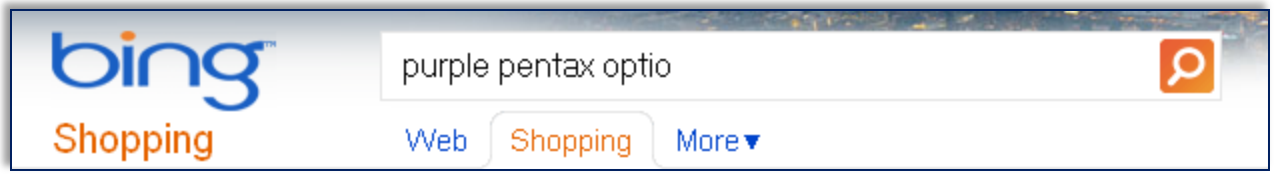
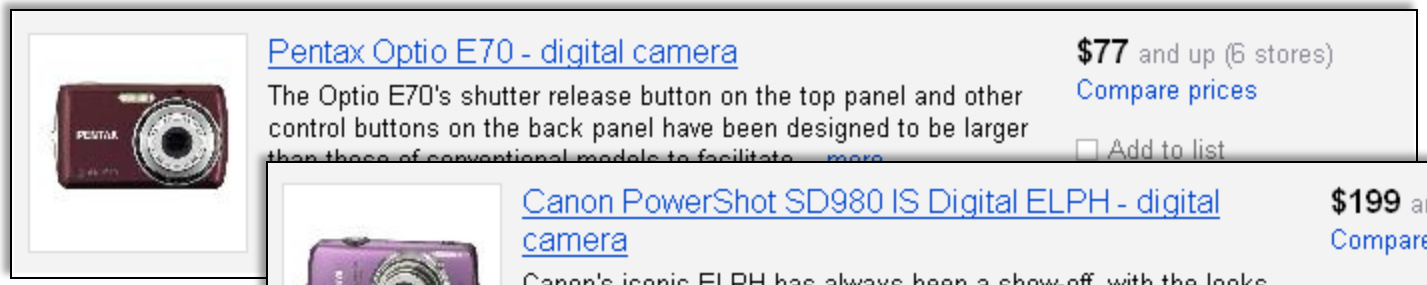
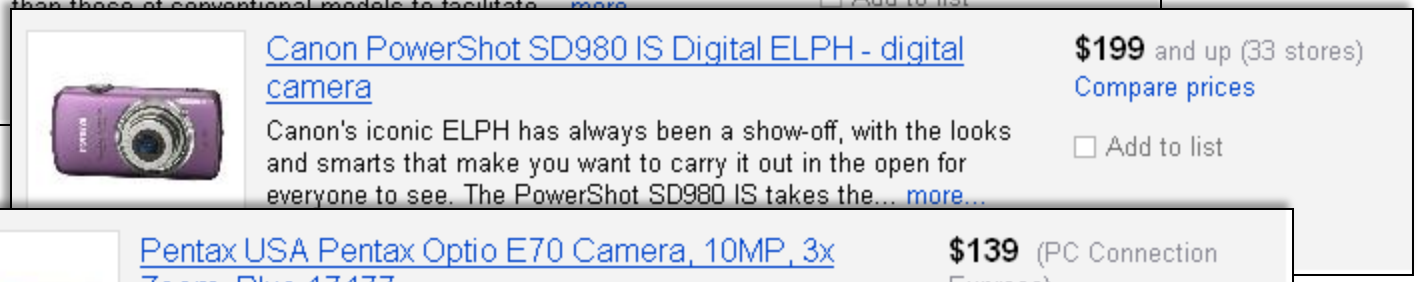
Result Enrichment in Commerce Search using Browse Trails

Debmalya Panigrahi
MIT

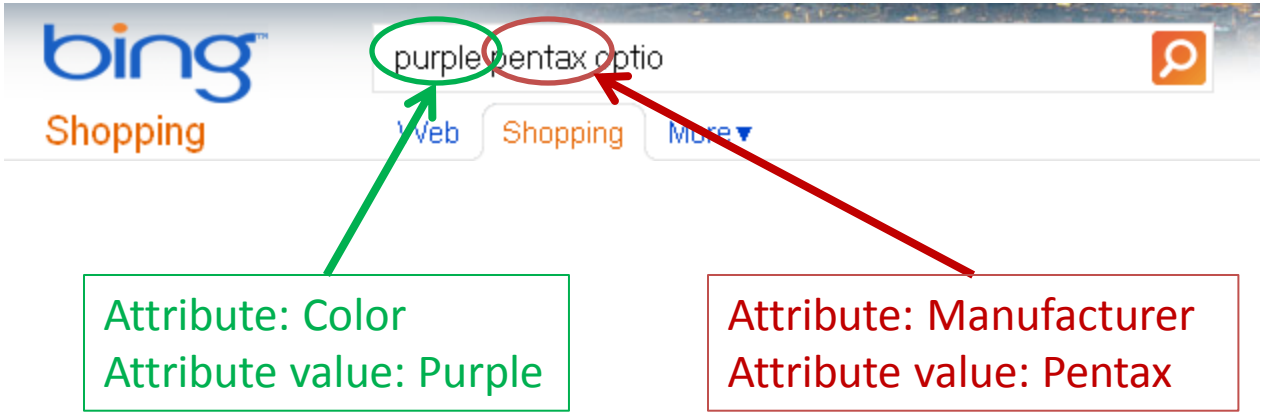
(Joint work with Sreenivas Gollapudi)


bing
Shopping
purple pentax optio
Web
Shopping
More


Pentax Optio E70 - digital camera
The Optio E70's shutter release button on the top panel and other control buttons on the back panel have been designed to be larger
$77 and up (6 stores)
Compare prices
Add to list


Canon PowerShot SD980 IS Digital ELPH - digital camera
Canon's iconic ELPH has always been a show-off, with the looks and smarts that make you want to carry it out in the open for everyone to see. The PowerShot SD980 IS takes the... more...
$199 and up (33 stores)
Compare prices
Add to list

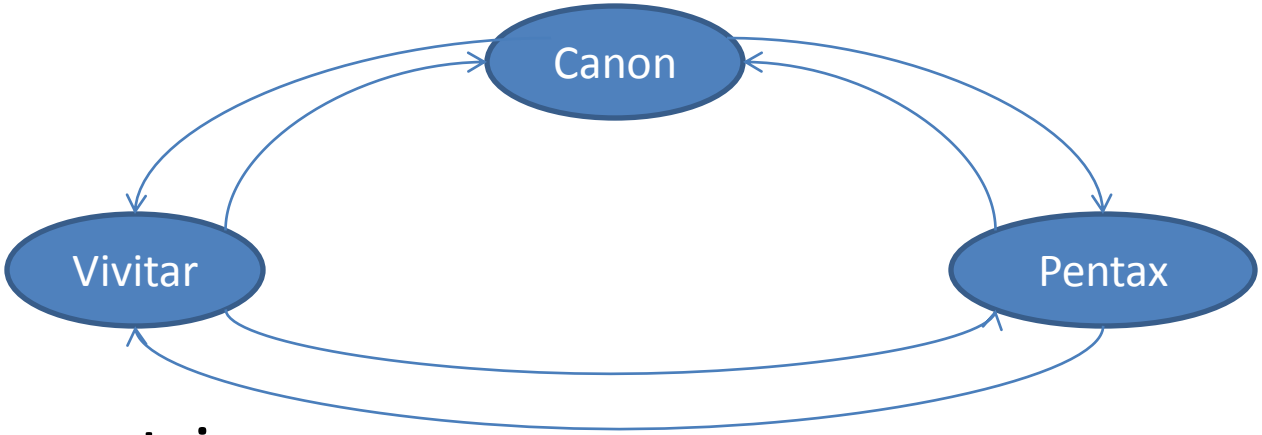

Pentax USA Pentax Optio E70 Camera, 10MP, 3x Zoom, Blue 17477
The Optio E70's shutter release button on the top panel and other control buttons on the back panel have been designed to be larger than those of conventional models to facilitate... more...
$139 (PC Connection Express)
Go to store
See store for details



- How important (non-replaceable) is an attribute value?
  - Replace purple or Pentax?
- What attribute value should be replace it with?
  - Replace purple with red or blue?

Similarity/Replaceability

- Sim(Pentax, Canon): Relative frequency of Pentax queries visiting web domains relevant to Canon
  - Sim(Pentax, Canon) >> Sim(Pentax, Vivitar)



- Asymmetric

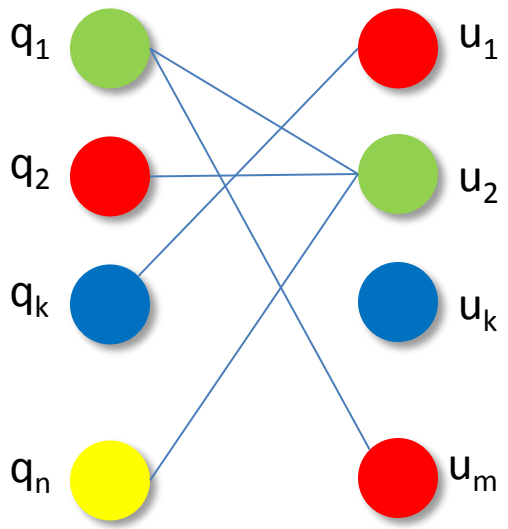
Importance/Non-replaceability

- Imp(Pentax): Relative frequency of Pentax queries visiting web domains relevant to Pentax
  - Imp(Pentax) = Sim(Pentax, Pentax)
  - Imp(Canon) >> Imp(Pentax) >> Imp(Vivitar)

- Key Question: What are the relevant attribute values for a web domain?

Previous Approaches

- Web crawling
  - Resource intensive
- Click-based methods
  - Spurious clicks
  - Sparse data (overcome by similarity estimations)



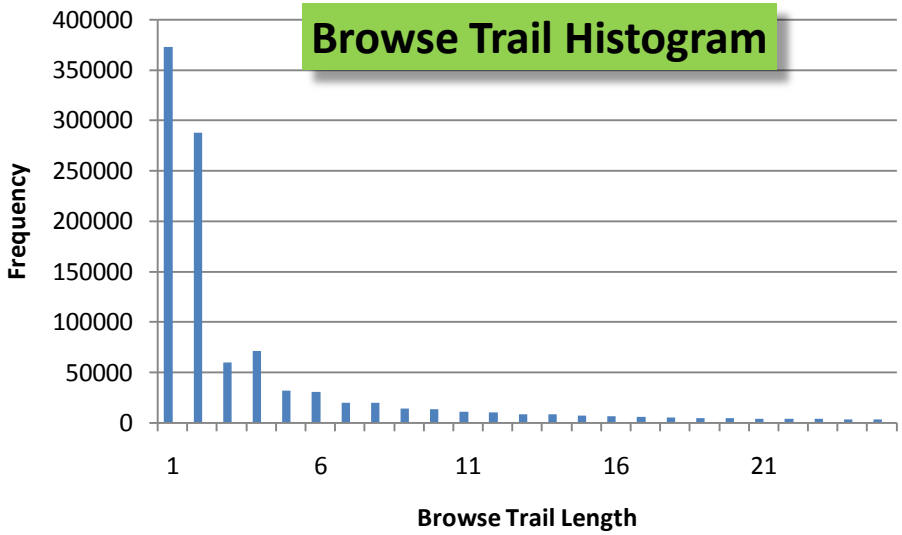
Query click graph

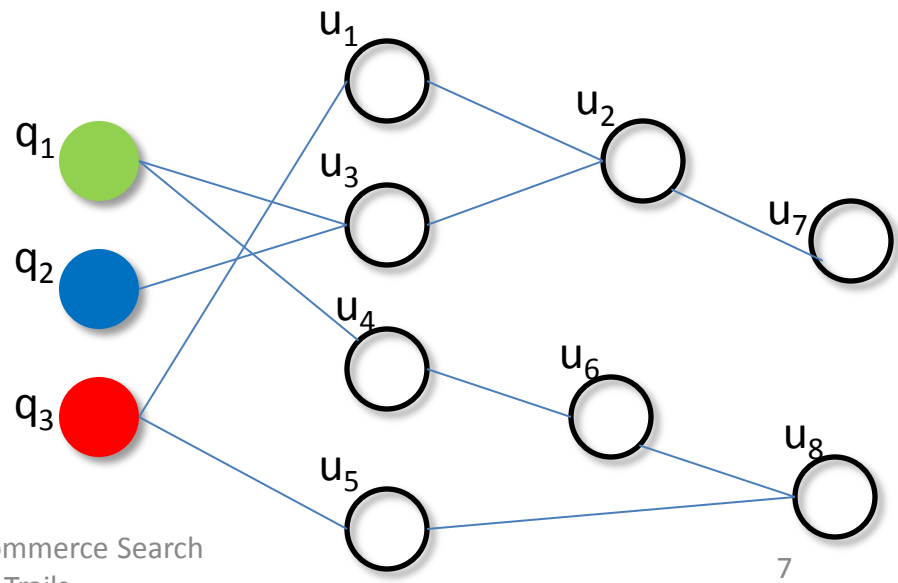
Beyond Search Clicks: Browse Trails

- Advantages
  - Larger corpus
  - More robust to noisy clicks
  - Indirect associations
- Previous usage
  - finding relevant search results

    [Bilenko-White '08]
  - finding popular destinations

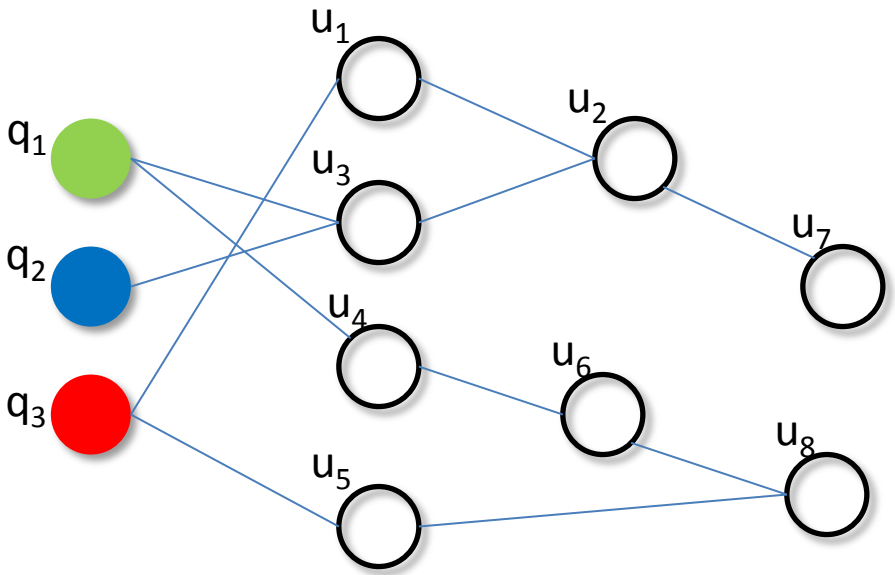
    [White *et al* '09]





Input Data: Browse Trails

- Why not assume more information about browsed pages?
  - Data source is browser: crawling pages huge overhead
  - Many pages in non-English languages: urls cannot be parsed to identify tokens
  - Humungous amount of data: mining of any additional information will incur huge overhead



Salient Features

- Category-based: homogenous set of queries
  - e.g., digital cameras
- Query comes annotated with values of important attributes
  - purple Pentax optio = purple + Pentax optio
- Each attribute treated separately

Rest of the Talk

- Annotating web domains (Algorithm + Analytical Framework)

- Experimental Results

Rest of the Talk

- Annotating web domains (Algorithm + Analytical Framework)

- Experimental Results

Annotating Web Domains

- For each web domain, create list of attribute values relevant to it based on frequency counts
  - Browse trails for queries containing Canon are more likely to visit canon.com than for Nikon

canon.com	amazon.com
Canon	Canon
Nikon	Nikon
Pentax	Pentax
…	…

(Simplified) Analytical Framework

- Suppose each domain d has a single (unknown) relevant attribute value k(d)
- f_{qd} = Frequency of browse trails for queries containing attribute value q visiting domain d

- $Sim(q, q') = \sum_{d:\, k(d) = q'} f_{qd} / \sum_d f_{qd}$
- $Imp(q) = \sum_{d:\, k(d) = q} f_{qd} / \sum_d f_{qd}$

- Input: Frequency counts f_{qd}
- Output: Values of Sim(q, q') and Imp(q) via identification of k(d)

Assumptions about input

- D_q: domains with k(d) = q
- Positive Bias: Queries with attribute value q visit domains in D_q more frequently than domains in any other $D_{q'}$

$$\sum_{d:\, k(d) = q} f_{qd} > \sum_{d:\, k(d) = q'} f_{qd}$$

- $n_{qq'} = \sum_{d:\, k(d) = q'} f_{qd}$ and $n_q = \sum_{q'} n_{qq'}$
- (α, δ)-Uniformity: Domains in D_q have similar query visit patterns

$$\text{If } d \in D_{q'} \text{ then } \alpha\,(n_{qq'} / |D_{q'}|) < f_{qd} < \delta\,(n_{qq'} / |D_{q'}|)$$

- (β, γ)-Proportionality: Relative proportion of an attribute value in queries and domains is similar

$$\beta\,(|D_q| / |D|) < n_q / \sum_{q'} n_{q'} < \gamma(|D_q| / |D|)$$

Annotation Result

- Theorem: There is an algorithm that annotates each domain d with a list of attribute values L_d such that the following properties are satisfied under the input assumptions:
  - Completeness:

    If $k(d) = q$ and $Imp(q) > \theta$, then $q \in L_d$
  - Soundness:

    If $k(d) = q$ and $q' \in L_d$, then $Sim(q, q') > (\alpha\beta\theta / \gamma\delta)^2$
  - Length:

    The average length of L_d is at most $(\gamma / \alpha\beta\theta)$

Rest of the Talk

- Annotating web domains (Algorithm + Analytical Framework)

- Experimental Results

Scalability

- Too many domains
  - Only retain frequently visited domains
- Too many attribute values per domain
  - Only retain frequently seen attribute values for the domain
- Implementation options on a stream of browse trails
  - “Heavy Hitters”
  - “Reservoir Sampling”

Experimental Validation of Assumptions

- Data: Use amazon.com pages annotated separately by parsing their url (ground truth)

- Positive Bias: Most queries with attribute value q mostly visited pages annotated by q
- Uniformity: The distribution of queries with attribute value q visiting pages annotated by q' had high entropy
- Proportionality: The relative frequency of attribute value q in queries and product pages was highly correlated

Experimental Data

- Dataset: IE 8 browse trail data from Feb-Jul 2009
  - 409 M browse trails
  - 63 % trails of length > 1
- Sample Categories: Digital cameras, Laptops, Kitchen Appliances, LCD TVs, …
  - Overall 26 top-level categories with 609 leaf-level categories
- Sample Attributes: Manufacturer, Product Line, Model, Screen Size, ..

Similarity Values

- Similarity scores computed by
  - Algorithm
  - Human Judges (using Amazon Mechanical Turk)

- For fixed q, ranks of attribute values q' in order of Sim(q, q'): $r_{Algo}(q, q')$ and $r_{Human}(q, q')$
- Agreement between lists:
  - $Disagree(q) = \sum_{q'} |1 / r_{Algo}(q, q') - 1 / r_{Human}(q, q')|$

Similarity Values

category	attribute	avg normalized $R(q)$
desktop computers	manufacturer	0.045
cabinet & drawer hardware	hardware material	0.063
mattresses	manufacturer	0.060
mowers & tractors	manufacturer	0.064
door hardware & locks	hardware material	0.070
cooktops	manufacturer	0.075
rings	stone	0.082
pants	bottom style	0.055
laptop computers	color	0.057
gardening tools	manufacturer	0.070
sweaters	apparel material	0.064
home theater systems	manufacturer	0.068
amplifiers	manufacturer	0.054
action figures	character	0.071
generators	manufacturer	0.051
vehicle playsets	manufacturer	0.065
vacuums	manufacturer	0.063
printers	manufacturer	0.055
radio controlled toys	manufacturer	0.079
cell phones	product line	0.065
cell phones	manufacturer	0.063
shirts	apparel material	0.066

Similarity Values

- Multiple attributes combined and treated as a single attribute

Product Query	Related Products		
sony vaio	sony vaio(0.550)	apple macbook (0.050)	hp pavilion (0.034)
lenovo thinkpad	lenovo thinkpad (0.157)	ibm thinkpad (0.130)	apple macbook (0.060)
pentax optio	panasonic lumix (0.219)	pentax optio (0.162)	canon powershot (0.151)
canon powershot	canon powershot (0.359)	panasonic lumix (0.160)	nikon coolpix (0.115)

Importance Values

Kitchen Appliances		Beds	Action Figures
Manufacturer	Color	Bed Type	Character
ge (0.23)	stainless steel (0.23)	Platform (0.23)	transformers (0.28)
lg (0.12)	black (0.06)	Bunk (0.19)	wwe (0.17)
samsung (0.12)	white (0.04)	Teen (0.17)	godzilla (0.14)
sharp (0.12)	steel (0.02)	Toddler (0.11)	rescue heroes (0.11)
maytag (0.10)	silver (0.02)	Loft (0.11)	spawn (0.09)
whirlpool (0.06)	gold (0.02)	Trundle (0.07)	star wars (0.05)
panasonic (0.05)	orange (0.01)	Kids (0.04)	halo (0.04)
electrolux (0.04)	clean steel (0.01)	Sleigh (0.03)	superman (0.02)
siemens (0.03)	graphite (0.01)	Canopy (0.03)	gundam (0.02)
frigidaire (0.03)	blue (0.01)	Adjustable (0.01)	lord of the rings (0.01)

Other applications

- Lists indicate which domains are good for which attribute values
  - Bestbuy may be good for Pentax but not for Canon cameras
- Product diversification
  - Interplay between importance and similarity

Future Work

- Dependencies among attributes
  - Few combinations, e.g. (manufacturer, model line) pairs): Treat attribute tuples as single attributes
  - How to handle many combinations, e.g. (manufacturer, color)?

- Questions?