

Detecting Online Commercial Intention (OCI)



Honghua (Kathy) Dai, Zaiqing Nie, Lee Wang, Lingzhi Zhao, Ji-Rong Wen, and Ying Li



Agenda

- Motivations and introduction to OCI (Online Commercial Intention)
- A machine learning-based approach for OCI detection
- Experiments
- Conclusion and future work



Motivation

- Serving ads will be more effective and less annoying, when user has intent to purchase
- We are interested in detecting web pages / queries that show intention to commit a commercial activity (purchase, rent, bid, or sell...)

OCI vs

3 search goal categories

- Navigational
 - The immediate intent is to reach a particular site
- Informational
 - The intent is to acquire some information assumed to be present on one or more web pages.
- Transactional
 - The intent is to perform some web-mediated activity

OCI can be seen as a new dimension of user search goals.

	Commercial	Non-Commercial
Navigational	walmart	hotmail
Informational	Digital camera	San Francisco
Transactional / Resource	U2 music download	Collide lyrics



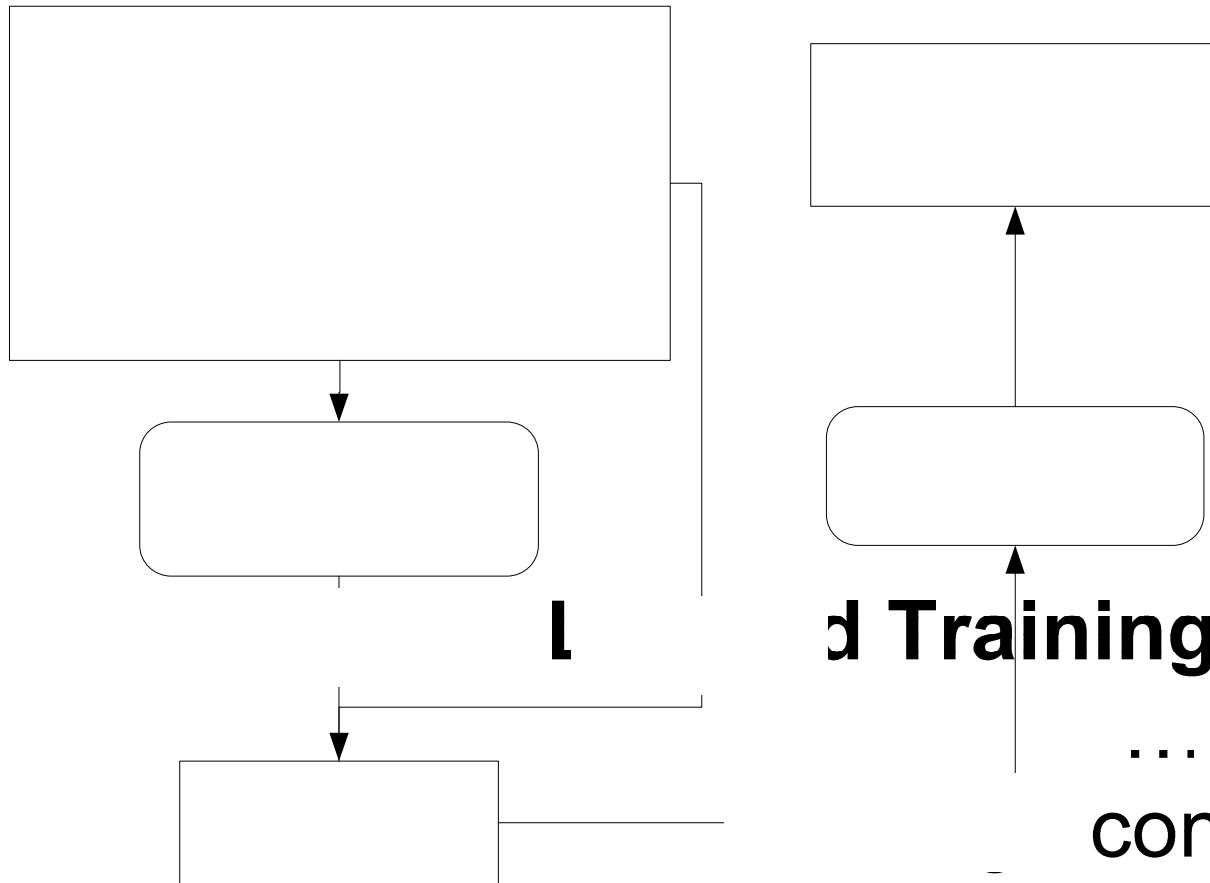
Define the OCI detection problem

- A binary classification problem
 - OCI: Query/Page \rightarrow {Commercial, Non-Commercial}
- We can derive the commercial sense from a confidence value that ranges from 0 (no commercial intent) to 1 (strong commercial intent)





Framework of Detecting Page OCI



d Training Page C

...

content of

<http://shopping.msn.com>

Commercial



Keywords selection

- Select significant and reliable keywords

- Significance:
$$Sig(k) = \frac{Max\{\Pr(k | C_+), \Pr(k | C_-)\}}{\Pr(k | C_-) + \Pr(k | C_+)} \times 2 - 1$$

- Frequency:
$$Freq(k) = \Pr(k | C_+ \cup C_-)$$

- Keyword selection threshold

- For simplicity we use the same threshold for the two measures in the experiments.



Page feature composition

- We define two aspects of properties for each keyword in a page p :
 - $nit(k_i, p)$ keyword occurrences in inner text
 - $nta(k_i, p)$ keyword occurrences in tag attributes
- As the result, a page p is represented by a feature vector using these two aspects



Detecting query OCI

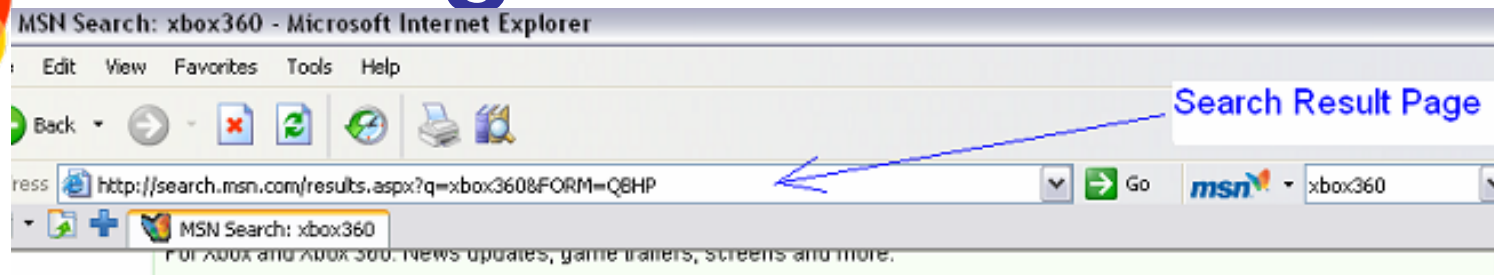
- Challenges

- Only few search queries contain explicit commercial indicators, such as “buy”, “price”, “rent”, “discount”, etc.
- Search queries are usually short.

- Solution

- Enrich query from external resource (search engine)
 - First result page (Query snippets)
 - Top N landing URLs
- Query classification problem -> page classification problem

Search result page and Landing URLs



[Xbox](#)

Microsoft's official guide to its Xbox gaming console offers press releases, machine specifications and visuals, a developer's list, and links.

www.xbox.com

[Xbox360 News, Views & Reviews](#)

Microsoft chairman Bill Gates has claimed that Microsoft will have shipped 10 million **Xbox360** units by the time the PS3 and Nintendo's Wii hit the shelves.

www.xbox360shop.net [Cached page](#)

[Xbox360](#)

Microsoft Offers **Xbox360** Video Downloads Epic has already dished out music videos for the Xbox 360 from the likes of Franz Ferdinand and Audioslave, and their catalogue also includes acts like Shakira ...

xbox-360er.blogspot.com [Cached page](#)

[xbox360](#)

xbox360 xbox360 look at, people wont if the logo is hard look at it The first few replies in a nintendo ds report being taunts of **xbox360** boy you question it, why then do we get

www.xbox360-emulators.com

[Deep Throat arrives for the XBox360](#)

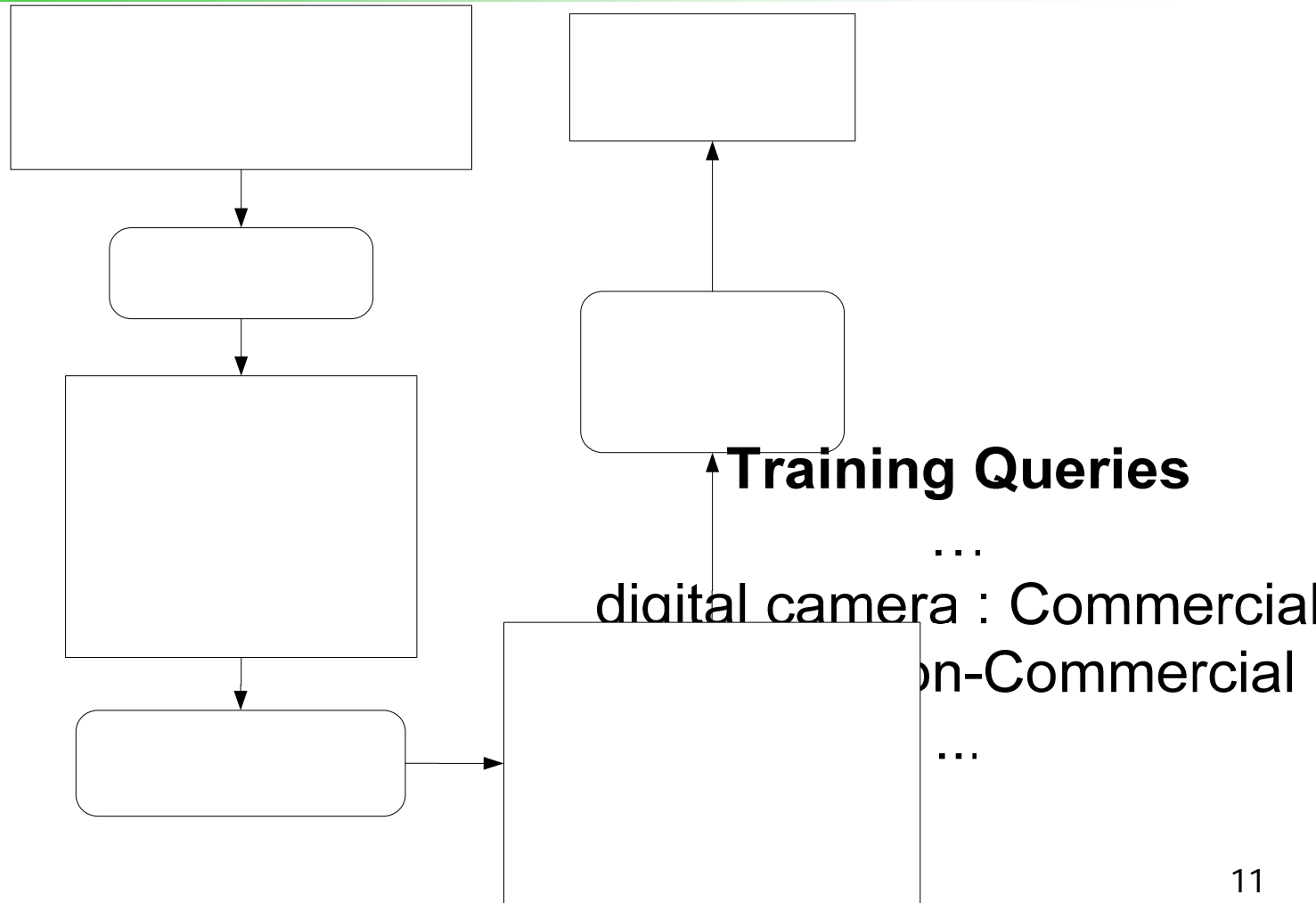
IF YOU THINK the **XBox360** lacked a killer app, well you are wrong, it has one, HD porn . If you think that it was just a fluke, check out HDHE, High Def Home Entertainment, a company the

www.theinquirer.net/?article=28837 [Cached page](#) 5/21/2006

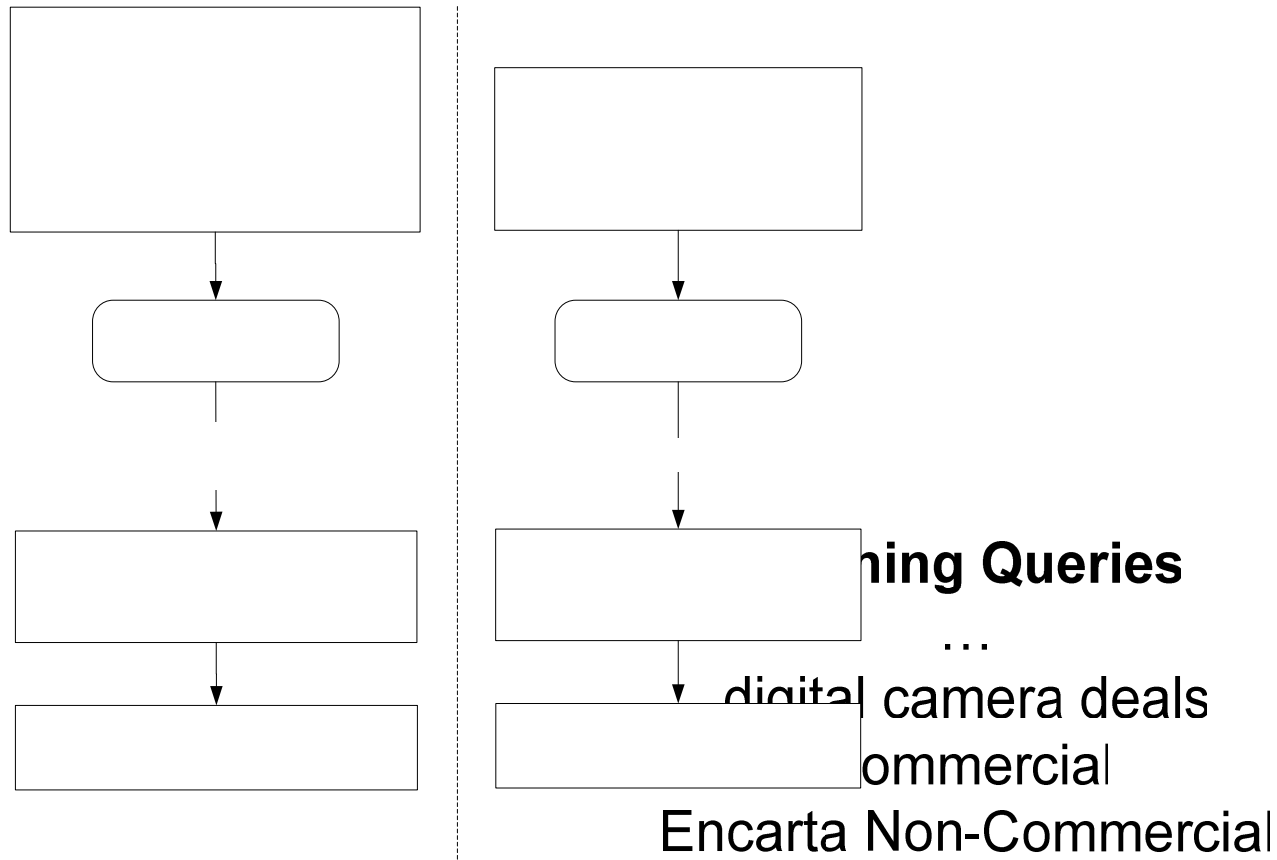
Query Snippet

Landing URL

Query OCI Detector based on Top N Landing URLs



Query OCI Detector based on first search result page



- Build a dedicated search result page classifier for this purpose



Labeling process

- We adopted majority vote: 3 human labelers voted for the labels
- Initial Web pages and queries were randomly selected from our page/query repository.

	<i>Pages</i>	<i>Queries</i>
Commercial	4074	602
Non-Commercial	21823	790
Total	25897	1408



Experiment Results - Page OCI detector

- Reach best performance (CF) when keyword selection threshold = 0.1 (using SVM as the classifier)

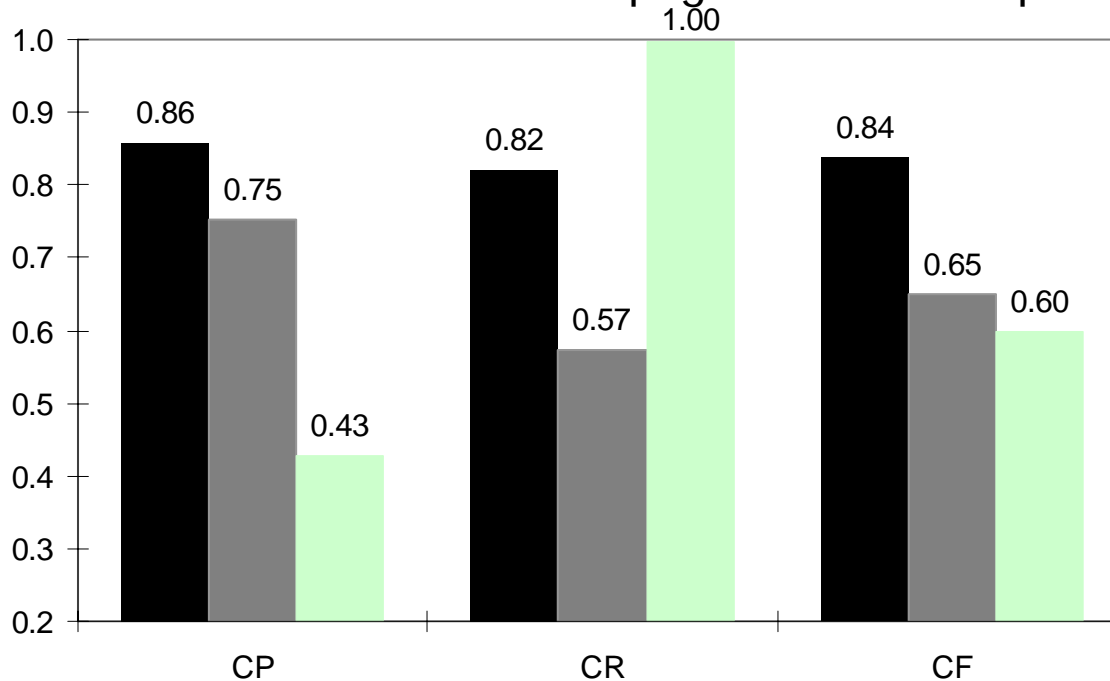
<i>Keyword Selection Threshold</i>	<i>Keyword Number</i>	<i>CP</i>	<i>CR</i>	<i>CF</i>
0.1	391	0.930	0.925	0.928

- CP, CR and CF are the precision, recall and F1 metrics for detecting commercial intent.



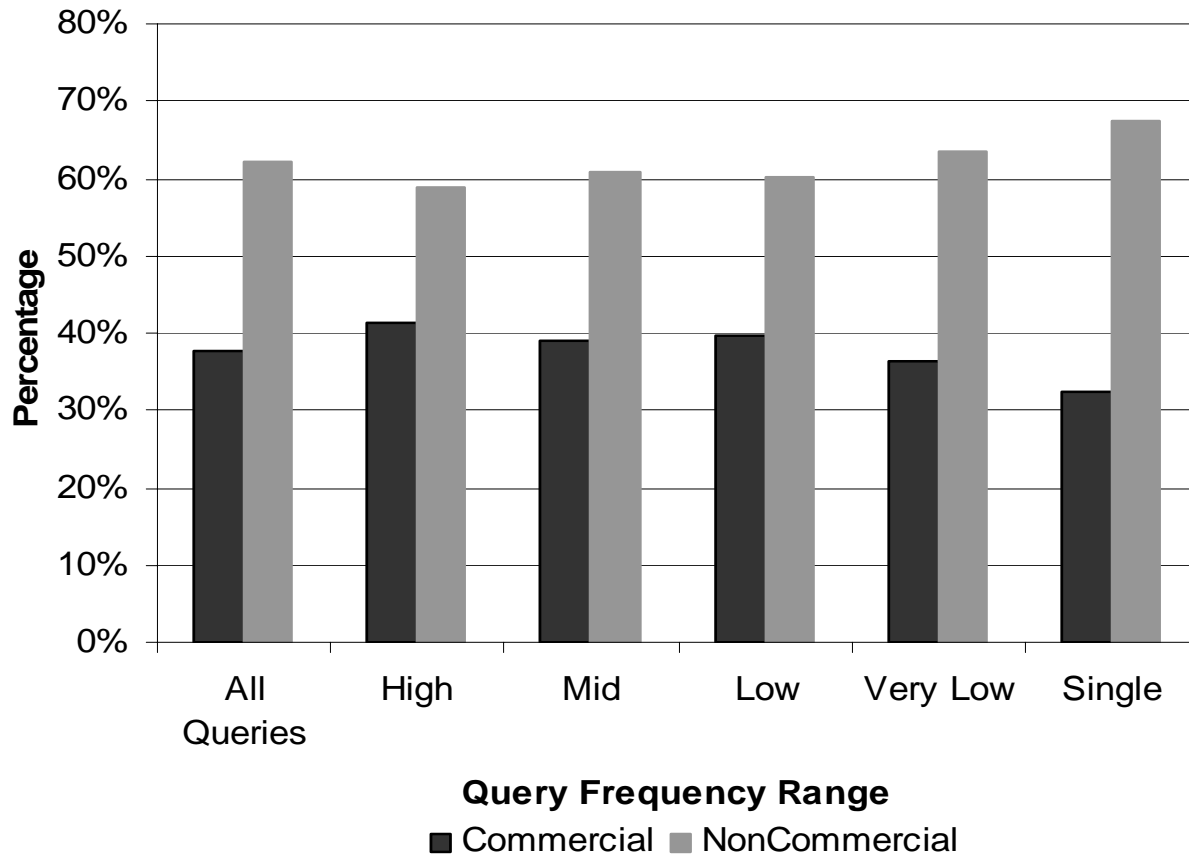
Experiment Results- Query OCI detector

■ Model based on first result page returns best performance.



- Model based on search result page: OCI(FSRPq)
- Model based on top N landing pages: OCI(TLPq) SVM
- Model based on top N landing pages: OCI(TLPq) Naïve Average

OCI Distribution among Query Frequency Ranges





Conclusions

- The notion of OCI (Online Commercial Intention) and the problem of detecting OCI from pages and queries.
- The framework of building machine learning models to detect OCI based on Web page content.
- Based on this framework, we build models to detect OCI from search queries.



Conclusions (cont.)

- Our framework trains learning models from two types of data sources for a given search query:
 - content of first search result page (query snippets)
 - content of top landing URLs returned by search engine.
- Experiments showed that the model based on the first search result page achieved better performance.
- We also discovered an interesting phenomenon that the portion of queries having commercial intention is higher in frequent query sets.



Future work

- Utilize search query click through logs
- Reduce labeling effort
- Take user online context into consideration in studying user's online intention
- Detect at which commercial activity phase a user is (research/commit).



Future Work (Cont.)

- Detect more detailed commercial intentions in different verticals
 - Traveling intention and preferences.
 - Branding awareness and preferences.
- Study how **specific** the user intention is:
 - “Halo2” vs “video games”
 - “cheap airline ticket new york to las vegas” vs “book a flight”
- Study the correlations between conversion rate and user intention.
- A lot of more interesting research problems!
 - We are HIRING!
 - Contact: KathyDai@microsoft.com



**Thank You for
Your Attention!**