Detecting Online Commercial Intention (OCI)

Honghua (Kathy) Dai
Microsoft Corporation
One Microsoft Way,
Redmond, WA 98052, USA
kathydai@microsoft.com

Zaiqing Nie
Microsoft Research Asia
Beijing, China
znie@microsoft.com

Lee Wang
leew2k@hotmail.com

Lingzhi Zhao
Tsinghua University
Beijing, China
zhaolz03@mails.tsinghua.edu.cn

Ji-Rong Wen
Microsoft Research Asia
Beijing, China
jrwen@microsoft.com

Ying Li
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052, USA
yingli@microsoft.com

ABSTRACT
Understanding goals and preferences behind a user's online activities can greatly help information providers, such as search engine and E-Commerce web sites, to personalize contents and thus improve user satisfaction. Understanding a user's intention could also provide other business advantages to information providers. For example, information providers can decide whether to display commercial content based on user's intent to purchase. Previous work on Web search defines three major types of user search goals for search queries: navigational, informational and transactional or resource [1][7]. In this paper, we focus our attention on capturing commercial intention from search queries and Web pages, i.e., when a user submits the query or browse a Web page, whether he / she is about to commit or in the middle of a commercial activity, such as purchase, auction, selling, paid service, etc. We call the commercial intentions behind a user's online activities as OCI (Online Commercial Intention). We also propose the notion of "Commercial Activity Phase" (CAP), which identifies in which phase a user is in his/her commercial activities: Research or Commit. We present the framework of building machine learning models to learn OCI based on any Web page content. Based on that framework, we build models to detect OCI from search queries and Web pages. We train machine learning models from two types of data sources for a given search query: content of algorithmic search result page(s) and contents of top sites returned by a search engine. Our experiments show that the model based on the first data source achieved better performance. We also discover that frequent queries are more likely to have commercial intention. Finally we propose our future work in learning richer commercial intention behind users' online activities.

Categories and Subject Descriptors
I.6.5 [**MODELS AND PRINCIPLES**]: Model Development – *Modeling methodologies.*

General Terms
Management, Measurement, Performance, Design, Experimentation, Human Factors.

Keywords
Intention, Search Intention, Online Commercial Intention, OCI, SVM.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

1. INTRODUCTION
There are two major online user activities on the Web. The first type of user activities is the well-studied browsing activity, i.e., how user visits Web pages on one or more Web sites. The second type, searching activity, is under great attention recently. Since the past decade, people started to study the phenomenon of search engines, their impact and user search behavior by surveying, statistical log analysis, and search presentation study [1][3][6][7][8]. A comprehensive review on Web searching studies can be found in [13]. Recently there has been more and more work being done in the field of understanding goals and intention of search users. Understanding goals and preferences behind user's search activities can help different types of information providers (e.g., search engines, E-Commerce sites, and online advertising businesses), to personalize search results and thus improve user satisfaction. Pilot research and applications can be found in [11][10][2][12] and [14] .

In [1] and [7], user's search intention / goals were classified into three general categories: Navigational, Informational and Transactional or Resource. The goal of a navigational query is to reach a particular web site; the intent of an informational query is to acquire information on web pages; and a user who inputs transactional queries are to perform some "web-mediated" activity. User search goals can also be represented using topical categories ([18][9] and [22]) or location attributes [21]. A few efforts have been invested in automatically identify user search goals [4][10][20][21].

Often times, information providers would like to know whether a user has intention to purchase or participate in commercial services, which we call "Online Commercial Intention" or **OCI**. Online Commercial Intention has broader scope than general search intention discussed in [1] and [7]. First of all, OCI can be applied on both searching and browsing activities. Secondly, OCI of a search user can be seen as another independent dimension of search intention besides the three categories: Navigational, Informational and Transactional / Resource. Table 1 shows that Online Commercial Intention (OCI) can cover all three of the

previously defined types of search goals and can be seen as a new dimension of user search goals.

Table 1: Online Commercial Intention vs Three Search Goal Categories

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

In addition, Online Commercial Intention can be further extended to reflect what phase a user is in during a user's commercial activity. We call it Commercial Activity Phase (CAP). Usually online users will do some research before making their mind to purchase. Therefore we define two Commercial Activity Phases: Research and Commit. During research phase, users may search for general information of their need, look at product information and reviews, search for deals, compare prices, etc. They may come to commit phase when they make upx their mind. Usually at this phase users will reach the transaction pages, or go offline to complete the commercial activity.

Knowing Online Commercial Intention (OCI), information providers could adopt different strategies when providing service to users who intend to purchase versus the users who don't intend to purchase. For example, commercial intention can be integrated into any recommender system's ranking algorithm so that recommendation results can be ranked based on the OCI. Another advantage of learning commercial intention is that information providers can detect user's commercial value. A user submits query "who is the 20th president of United States" has little commercial value, while the user who submits query "honeymoon suite in Maui" is much more attractive to Ecommerce Web sites, search engines or advertisers because they are more likely to be online shoppers.

When studying Online Commercial Intention (OCI), we can learn individual users' intention from their behavior; we can also understand the general intention at the level of search queries or Web pages. In this paper, we focus our attention on OCI of search queries or Web pages.

We define online commercial Intention (OCI) to be a function from a query or Web page to binary value: **Commercial** or **Non-Commercial**. More specifically, if the general purpose of users submitting a query or visiting a Web page is to commit a commercial activity, such as purchase, auction, selling, or **paid service**, the query / Web page will be treated as Commercial. If accessing the query or the Web page has little to do with any commercial service or activity, the query / Web page is considered as non-commercial.

There are several challenges in determining OCI from search queries. First of all, most search query terms do not explicitly contain terms that tell commercial intention. We found that only a small percentage of search queries have explicit terms indicating commercial intention. Furthermore, search queries are usually very short. According to [13] and [6], the average length of search queries in general purpose search engines is about two terms. Therefore, we will need help from external data sources in order to capture queries' online commercial intention.

In the last two years, there have been studies in learning different types of attributes from search queries by means of external information. In [10], authors adopted multiple types of information, e.g., different query term distribution in independent document set, mutual information, the usage rate as anchor texts, and POS information, to classify search query into navigational or informational categories. The same query categorization problem was visited again by Lee et al. [20]. They built query classifiers based on past user-click behavior and anchor-link distribution to achieve the same query categorization problem. Lee Wang et al. [21] built location information detector based on multiple data sources, including query result page content (snippets) and query logs. In [15], similarity between two queries was computed from both the keywords similarity and the common search result landing pages selected by users. For Web pages, the problem is less serious because pages are usually longer than search queries.

Subjectivity of the definition of OCI is another challenge for both search queries and Web pages. Some search queries / Web pages may be ambiguous in their senses of commercial intention. A user may be interested in learning the non-commercial side of the information about a commercial product/service. However, because we consider the general purpose of the query/Web page, i.e., the purpose that is agreed by majority of users who used this query, we assume that an average online user has a commercial intention if he/she accesses a query or a Web page about a commercial product or service. We will discuss the details of labeling process in 4.2.

In this paper, we present a solution to detect online commercial intention (OCI). The contributions of this paper include:

1. A formal definition of online commercial intention – **OCI**
2. A notion of Commercial Activity Phase (CAP)
3. A supervised learning system to learn OCI of search queries and Web pages
4. A comprehensive evaluation of our solutions.
5. An interesting finding about the relation between query frequency and query's OCI.

The rest of this paper is arranged as following: Section 2 introduces the definition of OCI and section 3 discusses our methodology of building machine learning models to detect OCI from search queries and Web pages. Section 4 presents experiment and evaluation. In section 5 we propose future work in learning the online commercial intention framework and related applications.

2. Defining OCI: Online Commercial Intention

We define online commercial Intention (OCI) to be a function from a query or a Web page to a binary value: **Commercial** or **Non-Commercial**. More specifically, if the general purpose of users submitting the query or visiting a Web page is to commit a commercial activity, such as purchase, auction, selling, or paid service, the query can be treated as Commercial. Otherwise, the query / Web page is considered as non-commercial. We treat the problem of determining OCI as binary classification: given a search query / Web page, assign the query / Web page into one of the two classes: Commercial versus Non-Commercial.

Here we formally define OCI (Online Commercial Intention).

Given:

1. **Terms**: T is the set of all possible terms.

2. **Queries**: Q is the set of search queries of which we want to determine the commercial intention. A search query q is a sequence of terms in T.

3. Domain of Web pages **P** as the set of all Web pages on the Web.

4. Online **Commercial Intention Value** = {Commercial, Non-Commercial}

Our goal is to compute two functions

$$OCI : Q \rightarrow \{\text{Commercial, Non - Commercial}\}$$

$$OCI : P \rightarrow \{\text{Commercial, Non - Commercial}\}$$

Let's also define $p_q^k \in P$ to be the landing page on rank k in the search result of query q.

3. Learning Online Commercial Intention

Here we introduce our framework to detect OCI from Web pages / search queries. Our framework takes the approach of extracting features from page content and building the classifiers based on those features.

One intuitive approach is to utilize existing concept hierarchy or categories. For an existing taxonomy, we could label each concept/category to "Commercial" or "Non-Commercial". Then we map page content to matching concepts or topic categories, The OCI label of the concept/category will become the OCI of the page. Let's call this approach "the taxonomy-based approach".

The taxonomy based approach has shortcomings. Firstly for many categories (taxonomy entries), it is very difficult to assign a single commercial tag or a non-commercial tag to them. For example, for the category "art", it contains both commercial and non-commercial Web pages: a page about art history is a Non-Commercial page; while a painting auction page is a Commercial page. Secondly, the taxonomy based approach is less efficient. The taxonomy based approach needs to extract features for each category entry, whereas if we want to separate the commercial/non-commercial pages, we just need to extract the commercial features to do the classification. Lastly, the taxonomy-based approach may be less accurate because the categorization error of all the entries will be aggregated with the commercial/non-commercial tagging error.

In the following subsections, we discuss the process of acquiring labeled data and learning OCI from Web pages and search queries.

3.1 Labeling Process

It is desirable if we could probe into the whole life cycle of user's online activity to understand the user's intention. Here is an ideal labeling scenario: assume we have complete user activity histories, including every action from the beginning to the end of the user session. If the session leads to a completion of a commercial activity, the queries and Web pages in this session should be labeled as commercial. When there is a statistically sufficient number of such sessions for a specific query or Web page, we would be able to label the query or Web page's commercial intention using the majority vote: if majority of sessions containing this query / Web page led to a commercial activity, label the query / Web page as commercial.

In reality it is hard and expensive to collect the data described above. Many online sessions with commercial intention may not end up with online commercial activity. After acquiring price and other product information, many search users may call the stores or go to the stores physically to complete the purchase. Furthermore, for detecting OCI of search queries, tracking user activity after user leaves the search engine would require client side agent, which brings privacy issues.

We adopted the human-evaluation approach: asking human labelers to judge the general commercial intention of search queries or Web pages. In the labeling process, we asked human labelers to consider the purpose of the search query or Web pages from the perspective of general online users. If the general purpose of submitting the query or visiting the page is to commit a commercial activity, human labeler should label it as "Commercial". Human labelers will mark the query or the Web page as "Non-Commercial" if the general purpose of submitting the query or visiting the page has little to do with any commercial activity. If a human labeler is not sure about the intention behind the query / Web page, she/he can label it as "Confusing". A page is labeled only if majority of labelers agree on the labeling. Note that each search query / Web page is labeled with its **general** commercial intention. In this paper, we do not consider the commercial intention in individual user sessions.

3.2 Web Page OCI Detector

In this section, we will build a framework to learn Web page OCI. We call the model built from this framework as "Page OCI Detector". More specifically, given a Web page P, the page OCI detector is designed to detect OCI of p: $OCI(p)$.

The framework is described in the following figure:

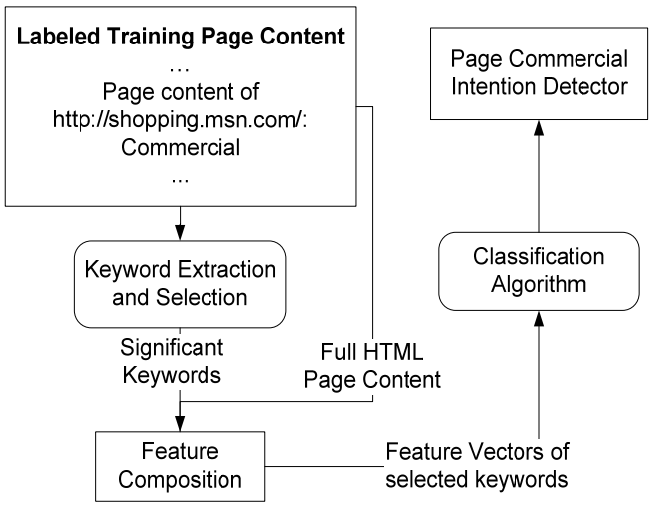

Figure 1: Framework of Learning Page OCI - Training

We first extract keywords (here a keyword is a single term in T) from both inner text and tag attributes of all the labeled Web page in the training data.

The next step is feature selection. Intuitively a good feature should be "significant" in order to distinguish between class labels. At the same time, it should also be frequent enough to be reliable and representative.

The measure of significance is defined as following:

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

and the measure of frequency is defined as following:

$$Freq(k) = \Pr(k \mid C_+ \cup C_-)$$

where $\Pr(k \mid C)$ is the probability of the keyword k occurring in a Web page belonging to class C. Here C_+ means positive class and C_- represents negative class (commercial and non-commercial respectively in our case). Thus, $Sig(k)$ and $Freq(k)$ are real numbers between 0 and 1. We set thresholds for $Sig(k)$ and $Freq(k)$ to select "good" keywords.

After keyword extraction and selection, we will obtain a keyword set $K = \{k_1, k_2, \cdots, k_n\}$ where k_i is the ith selected keyword and n is the total number of selected keywords.

We define two aspects of properties for each keyword k_i in a page p:

1. $nit(k_i, p)$ = Number of elements that the keyword appeared in its inner text in p / Total number of elements in page p.
2. $nta(k_i, p)$ = Number of elements that the keyword appeared in its tag attributes in p / Total number of elements in page p.

The first aspect reflects the general textual information of the Web page, while the second dimension tracks the text on special elements, such as buttons, images, and forms, etc. For example, text "order" on a button has very different role of its appearance as general text. Both numbers were also smoothed by power of 1/8 because most of them are too close to 0.

Thus, for a page p with n keywords, page p can be represented using a vector with 2*n dimensions:

$$kv_p = \langle nit(k_1, p), nta(k_1, p), \cdots, nit(k_n, p), nta(k_n, p) \rangle$$

Such vectors will become the input to a standard classification algorithm. In our experiment, we adopt SVM [17] as the classification algorithm. Using SVM algorithm, we acquire a model for page commercial intention detection.

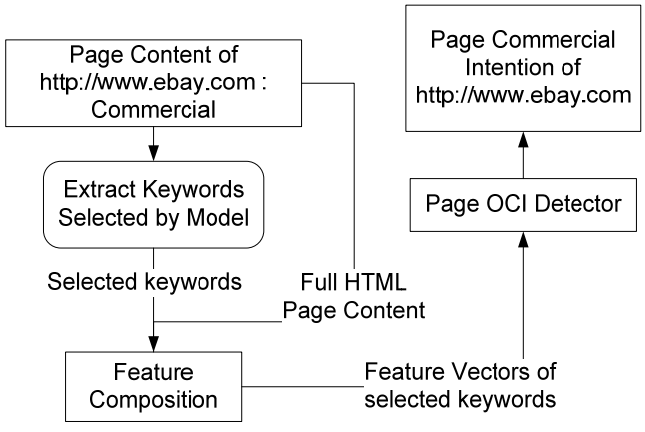

Figure 2: Framework of Learning Page OCI – Prediction

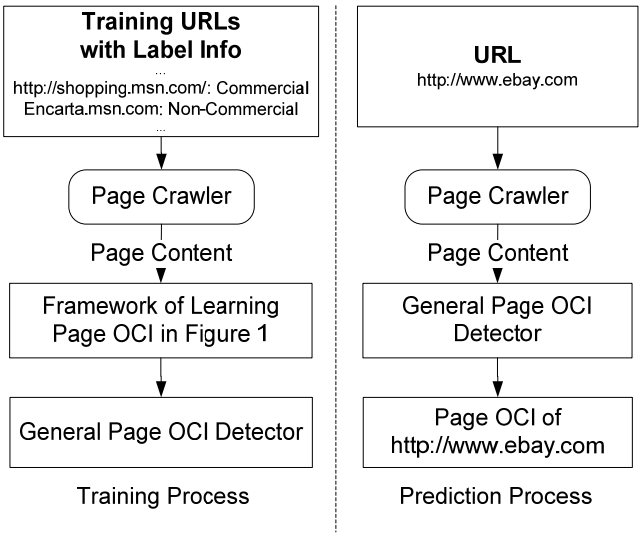

Figure 3: Build a General Page OCI Detector

When predicting a Web page's commercial intention, we will need to extract the features via the keywords that have been selected in the training phase from the target Web page. Then we use the same 2-aspect approach to form a 2*n dimensional input vector to the model built in the training phase. Figure 2 shows the process of detecting commercial intention based on page content. Figure 3 depicts the process of training the general Web page OCI detector and using it to make predictions. The details of labeling process and analysis are discussed in sections 3.1 and 4.2.

3.3 Query OCI Detector

3.3.1 Data Sources

There are four types of data sources that are available and may contribute to detecting the query OCI:

1. Constituent terms of search query
2. Content of top landing pages recommended by search engine
3. Content of search result page
4. The number of user clicks of landing pages recommended by search engine

Some queries can be detected with their commercial intentions directly because they have explicit commercial indicators in the queries, such as "airline ticket **deals**" or "digital camera **price**". However, most other queries with commercial intention do not contain explicit commercial indictors, e.g., "used car", or "home depot". In our initial investigation, we composed a set of explicit commercial indicators based on common sense and knowledge. These terms include: "price", "cheap", "buy", "sell", "sale", "rent", "purchase", "auction", "deal", "coupon", "discount", "lease", "bargain", "retail", "advertise", "bidding", and "market" etc. Our statistics show that a very small percentage of queries contain explicit commercial intention indicators. Furthermore, it is hard to find explicit indicators for non-commercial intentions. Relying on search query terms will also suffer badly from the problem of data sparseness.

Landing pages recommended by search engine (the second data source), especially at top ranks, give deeper exploration of a

user's intention when we study the actual page content of these top-returned result URLs.

Search result pages (the third data source) usually contain title, short descriptions, and URL links to the recommended landing pages. The title and description provide relevant information about what the user is searching for. MSN search engine also provides relatively constant length for titles and descriptions in search result page. Therefore the length of search result page is much more stable comparing to arbitrary Web pages and search queries. Nowadays search engines also return sponsored links to commercial queries. Inclusion of sponsored links is helpful in detecting commercial search intention. In this paper, we are interested in the **first** search result page because it usually contains the most relevant result for search queries.

In addition, if we could observe the click distribution of landing pages recommended by search engine (the fourth data source), we would be able to get a good understanding of the general purpose of the query by applying statistical analysis on how users select search results. However, users may have certain level of trust on the results provided by search engine. The levels of trust may generate some bias on which result URLs a user picks.

We propose to build OCI detectors on the 2nd and the 3rd data sources, that is, content of top landing pages and content of first search result page. We noticed that both the 2nd and the 3rd data sources are based on Web page content: search result page by itself is a Web page, and top landing URLs recommended by search engine are a collection of Web pages. Therefore we can reuse the framework of learning Web page OCI to learn commercial intention from search result page returned by search engine. The framework enables us to build one model for each type of Web page.

We are particularly interested in finding out which data source is more effective: query snippets, result page content, or the combination of the two data sources.

3.3.2 Detecting OCI based on Top Search Result Landing Pages

Most search queries do not contain explicit indicators for commercial intentions. A typical example is "digital camera". Therefore, we will need external help to understand the meaning of the query.

In this section we will build models using the content of Web pages that have top rankings in search results to predict query commercial intention. There are usually a constant number of recommended landing pages (the number is usually 10 in MSN search) in the first result page returned from search engine.

Note that we already have a general page OCI detector (described in 3.2). We just need to use the detector to detect the OCI of top search result landing pages, and then combine the results of page OCI together as the query OCI. For a query q, sending the top N search result landing pages to general page OCI detector will get an N dimensional vector:

$$\langle OCI(p_q^1), OCI(p_q^2), \cdots, OCI(p_q^N) \rangle$$

where p_q^i is the Web page that has rank i in the search result of query q.

In order to compute the query OCI, the simplest method is to average the OCI of the top N search result landing pages. However, this method ignores the factor of the ranking of search results. For example, higher ranking landing pages are likely to be more relevant to the query. In order to consider rankings in combining the page OCI of top URLs, we let the supervised learning algorithm to learn the best combination of the N factors. We use SVM to train the combining factors. The input of the SVM algorithm is the N dimensional vector above. We use the notion of $OCI(TLP_q)$ to represent the OCI predicted from the top landing pages for query q.

Figure 4 describes the logic flow of training the models based on top N search result landing pages. Figure 5 describes the logic flow of predicting query OCI based on the models from top N search result landing pages.

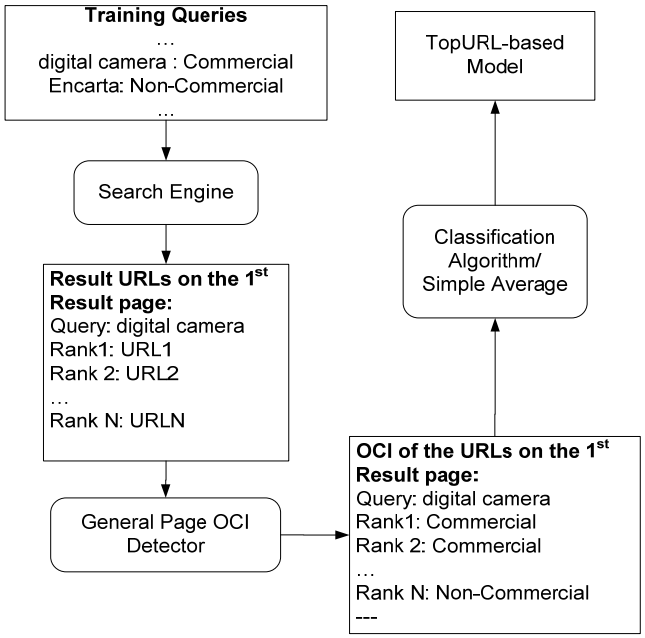

Figure 4: Detect OCI based on Search Result Landing Pages – Training

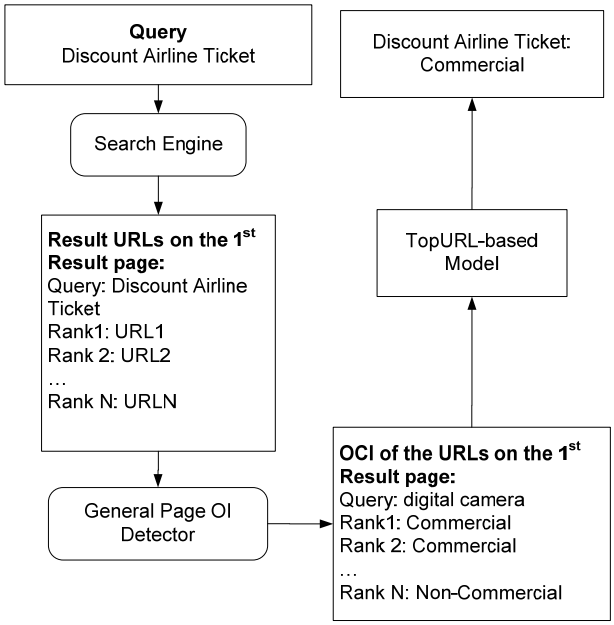

Figure 5: Detect OCI based on Search Result Landing Pages - Prediction

3.3.3 Detecting OCI based on First Search Result Page

First search result pages usually contain structured information on each recommended landing pages. We call such structured information "Query Snippets". They are valuable because:

1) Search engine provide reasonable relevance in the first result page;

2) Query snippets usually contain the part of the landing pages that matches the search query, therefore, more relevant than rest of the page content;

3) Query snippets usually have controlled length, therefore convenient for text processing.

We use the notion of $OCI(FSRP_q)$ to represent the OCI predicted from the first search result page-based model for query q. Figure 6 illustrates the process of detecting query's OCI from first search result pages.

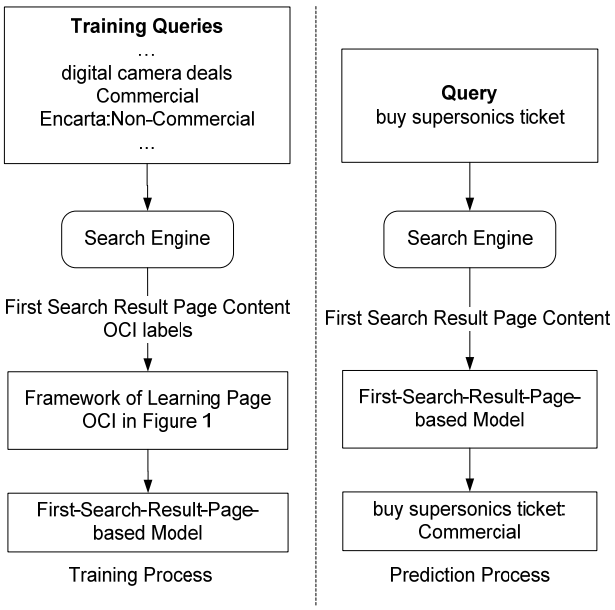

Figure 6: Detecting OCI from First Search Result Page

We chose to get search result pages from MSN search and use the default settings that give us 10 query snippets. For each training search query with its commercial intention label, we acquire the first search result page via the search engine. Thus the labeled query set will become the labeled first search result page set. And then we can apply the framework of learning page commercial intention (see Figure 1) on this labeled page set to build a search result page commercial intention detector.

When making a prediction on a search query's commercial intention, we first send the query to the search engine and acquire the first search result page. Applying the page to the search result page commercial intention detector will get the query's commercial intention.

4. Experiment

4.1 Data and Settings

There are a few data sets that we collected for the purpose of training and evaluation.

1. For the models for detecting query commercial intention, we randomly picked 1408 US English search queries from a one day MSN search log.
2. We collected the first search result page for the above 1408 search queries.
3. We also collected top 10 search result landing pages for the1408 search queries.
4. In order to build the general page commercial intention detector, we randomly picked 26186 English Web pages on the Web.

4.2 Labeling Analysis

We asked 3 human labelers to label the search queries and pages. Each query or page is labeled as "commercial", "non-commercial", or "confused". Each query was labeled by the 3 labelers separately. After labeling process on queries, we keep the queries / pages that were agreed by at least two labelers with non-confused labels. Finally we got 1392 "commercial" and "non-commercial" search queries and 25897 "commercial" and "non-commercial" English Web pages. Table 2 shows the distribution of pages and queries among labeled pages and queries.

Table 2: OCI Distribution among Labeled Pages and Queries

	Pages	*Queries*
Commercial	4074	602
Non-Commercial	21823	790
Confused	289	16
Total	26186	1408

This is an unbalanced classification problem, and the majority of Web pages are non-commercial pages. Since we could only use a limited number of training pages to ensure an acceptable convergence rate of the SVM training algorithm, we use an active learning approach to select some pages around the hyperplane as training examples. We selected all commercial pages and the equal number of non-commercial pages to train and test our initial model. After that, we asked the labelers to randomly pick pages on the Web and use the initial model to get the labels of those pages. We kept adding the misclassified pages into our data set. Thus we obtain a balanced data set containing 10964 pages. In the following section, we will use this data set to train and evaluate the page OCI detector. In this way, we could avoid including many non-commercial pages which are far away from the hyperplane. It's well known that those training examples that lie far away from the hyperplane barely participate in finding the hyperplane. Note that the mislabeled pages are added to both the training set and the testing set.

4.3 Evaluation Methodology

To evaluate the page OCI detector, we divided the data set prepared in the previous section into train set and test set equally.

Table 3 shows the distribution of commercial and non-commercial pages in training and test sets

Table 3: OCI Distribution of Experiment Page Data Sets

	Train Set	*Test Set*	*Total*
Commercial	2820	2936	5756
Non-Commercial	2555	2653	5208
Total	5375	5589	10964

We first extract and select keywords on the train set, and then train page OCI detector model based on train set. Finally, the model will be evaluated on the independent test set. In our experiments, we set the thresholds for $Sig(k)$ and $Freq(k)$ to select different number of keywords to see the performance based on different keyword number, and find the best threshold value.

To evaluate query OCI detector, we are interested in discovering what data source performs best in detecting OCI. We compare the model based on first search result page and the model based top N result landing pages using 3-fold cross validation. First, we randomly divide the 1392 queries into 3 folds, and then train a model on 2 folds and evaluate it on the other one.

We are interested in evaluating the detection power on commercial pages or queries by using standard IR evaluation measures [5] to evaluate our models:

$$Commercial\ \ precision(CP) = \frac{\#of\ \ pages/queries\ correctly\ classified\ as\ Commercial}{\#of\ \ pages/queries\ classified\ as\ Commercial}$$

$$Commercial\ \ recall(CR) = \frac{\#of\ \ pages/queries\ correctly\ classified\ as\ Commercial}{\#of\ \ pages/queries\ labeled\ as\ Commercial}$$

$$Commercial\ \ F1(CF) = \frac{2 \times CP \times CR}{CP + CR}$$

4.4 Results

4.4.1 Evaluating Page OCI Detector

In our experiments, the measure of keyword significance and frequency share the same threshold for simplicity. We call it keyword selection threshold.

Table 4: Performance of the Page OCI Detector

Kwd Selection Threshold	*Keyword Num.*	*CP*	*CR*	*CF*
0.01	4523	0.814	0.907	0.858
0.03	1712	0.956	0.884	0.919
0.05	989	0.948	0.899	0.923
0.075	600	0.934	0.918	0.926
0.1	**391**	**0.930**	**0.925**	**0.928**
0.15	179	0.921	0.923	0.922
0.2	100	0.916	0.905	0.910
0.3	25	0.893	0.840	0.865
0.4	6	0.848	0.791	0.819

The evaluation results for different keyword selection thresholds were shown in Table 4**Error! Reference source not found.**, Figure 7 and Figure 8. From Table 4**Error! Reference source not found.** and Figure 8, we get best performance in terms of CF when the threshold is 0.1. As we expected, the keyword number drops down sharply when the threshold increases (see Figure 7). And the model performance in terms of CF reaches a plateau to 0.928 when keyword selection threshold = 0.1 and then it gradually drops as when lifting the threshold.

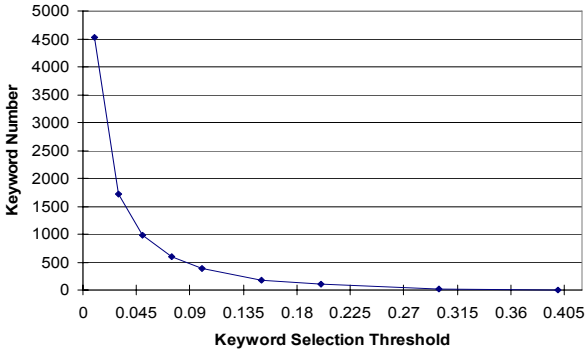

Figure 7: Number of Keywords Selected under Different Keyword Selection Thresholds

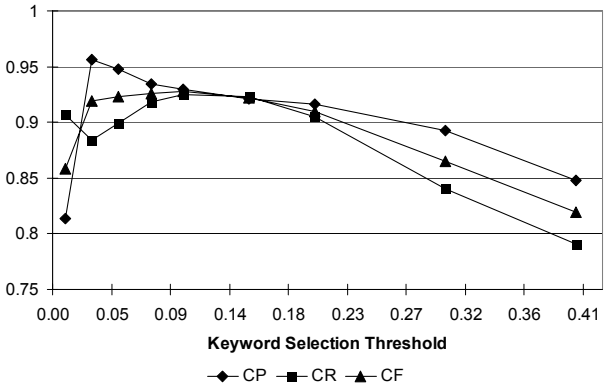

Figure 8: Page OCI Detector Model Performance under Different Keyword Selection Thresholds

4.4.2 Evaluating Query OCI Detector

In the experiments, we built models based on two types of data sources: content of search result page $OCI(FSRP_q)$, and content of top N landing pages $OCI(TLP_q)$. Recall from 3.3.2 that in order to compute $OCI(TLP_q)$, we first send top N search result landing pages to page OCI detector, then we combine the results from page OCI detector to compute the OCI of query q. For the reason of comparison, we build two models, one using SVM to train the combination factors, the other using naïve average to combine the results from page OCI detectors. The evaluation results are shown in the Figure 9. We can see the model based on first search result page has an obvious advantage comparing with the models based on top N result landing pages. The model using naïve combination has the worst performance. This can be explained from the fact that query snippets usually contain the most relevant information of Web pages that matches the query.

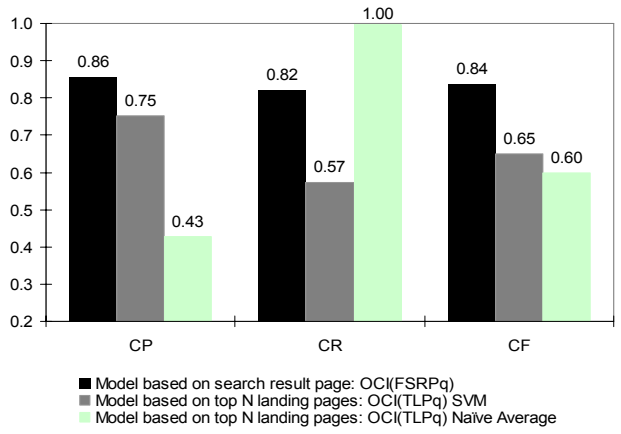

Figure 9: Query OCI Detector Performance

Comparison between 3 Models (3-fold Cross Validation)

4.4.3 OCI Analysis for a Stratified Query Sample based on Query Frequency

To gain deeper understanding of a general purpose search engine, we obtained a month search log from MSN search. We divided query frequency into 5 ranges: Single, Very Low, Low, Mid, and High. Queries falling into range "Single" are those queries only submitted once during one month period; while "High" frequency queries are the most popular queries. We randomly extract 10,000 queries in each query frequency range and form a 50,000 stratified sample.

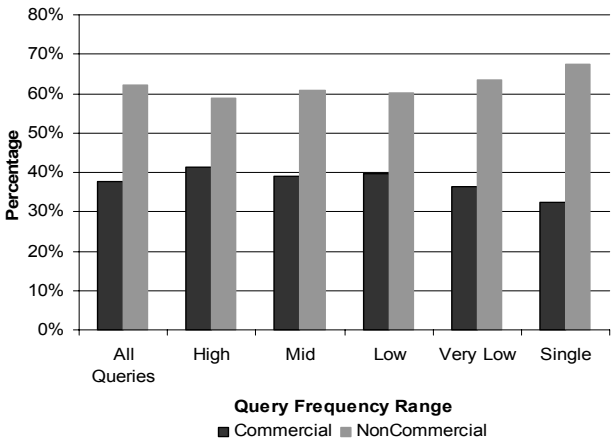

Figure 10: OCI Distribution among Query Frequency Ranges

We apply the query OCI detector based on first search result page on the 50000 query sample. Figure 10 shows the OCI distribution among the 50K queries, from which we can see that 38% of search queries have commercial intention. There is an interesting trend in search queries: query set with higher frequency usually have larger portion of queries with commercial intention. 32% of queries with single frequency have commercial intention, while the portion of queries with commercial intention is 41% in high frequency queries.

5. Conclusion and Future Work

The goal of this paper is to detect commercial intention from search queries and Web pages, i.e., when a user submits a query or browses a Web page, whether he / she is about to commit or is in the middle of a commercial activity, such as purchase, auction, selling, paid service, etc. We call the commercial intentions behind user's online activities OCI (Online Commercial Intention).

We also proposed the notion of "Commercial Activity Phase" (CAP), which identifies in which phase a user is in his/her commercial activities: Research or Commit.

We present the framework of building machine learning models to learn OCI based on any Web page content. Based on that framework, we build models to detect OCI from search queries and Web pages. Our framework trains learning models from two types of data sources for a given search query: content of returned first search result page and content of top pages returned by search engine. Our experiment showed that the model based on the first data source, i.e., returned first search result page content, achieved better performance.

We also discovered an interesting phenomenon that frequent queries are more likely to have commercial intention.

We will continue improving our algorithm to address the following issues:

1. Labeling effort. We are working on reducing labeling cost and subjectivity to improve the performance of the commercial intention detector. How to reduce the human subjectivity and automate labeling process is our next immediate task. We are interested in semi-supervised learning techniques to exploit the massive amount of unlabeled data.
2. Finding more effective and efficient features. In the future we will consider using other sources of data such as click data if it is available for the query.
3. Improving performance. Current solution requires sending queries to search engine, which make the unit response time rely on the search engine's response time. Solutions based on search result landing pages costs more time due to the cost from crawling top N pages. We will look into a more efficient method to obtain the context of search queries. We will also work on a better solution to find explicit commercial indicators, such as phrase-based solution.
4. Detecting Commercial Activity Phase (CAP). Current solution can detect whether or not a user has intention to commit a commercial activity. However, if we could tell at what stage the user is in the commercial activity, and how far it is to reach the actual purchase or other commitment, we could provide better personalized information service to individual users.
5. Detecting individual user's OCI based on his/her online behavior throughout an online session. Our current effort is focused on detecting general OCI for search queries or Web pages. However, as we have mentioned earlier, different users may have different intentions when accessing the same search query or Web pages, or may have different intentions on the same query in different phase of a purchase process. It will be interesting to consider the user session or online history as a whole unit to evaluate the user's OCI. It will be also worthwhile to investigate the evolution of a user's OCI over time.

ACKNOWLEDGMENTS

Our thanks to MSN [16] for allowing us to access the search data.

6. REFERENCES

[1] Andrei Broder. A taxonomy of web search. SIGIR Forum 36(2), 2002

[2] Anick, P. 2003. Using terminological feedback for web search refinement: a log-based study. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM Press, New York, NY, 88-95.

[3] Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., and Frieder, O. 2004. Hourly analysis of a very large topically categorized web query log. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Sheffield, United Kingdom, July 25 - 29, 2004). SIGIR '04. ACM Press, New York, NY, 321-328.

[4] Beitzel, S. M., Jensen, E. C., Frieder, O., Grossman, D., Lewis, D. D., Chowdhury, A., and Kolcz, A. 2005. Automatic web query classification using labeled and unlabeled training data. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05. ACM Press, New York, NY, 581-582.

[5] C. J. van Rijsbergen, Information Retrieval (Second Edition). London, U.K., 1979

[6] C. Silverstein, M. Henzinger, H. Marais, and M. Moricz. Analysis of a very large AltaVista query log. Technical Report 1998-014, Digital SRC, 1998.

[7] Rose, D.E. and Levinson, D.: Understanding user goals in web search. In Proceedings of the 13th international conference on World Wide Web, pages 13--19, ACM Press, 2004..

[8] Feng Qiu, Zhenyu Liu, Junghoo Cho "Analysis of User Web Traffic with a Focus on Search Activities." In Proceedings of the International Workshop on the Web and Databases (WebDB), June 2005.

[9] Google Directory. http://www.google.com/dirhp

[10] I. Kang and G. Kim. Query type classification for web document retrieval. In Proceedings of ACM SIGIR '03, 2003.

[11] Jaime Teevan, Susan T. Dumais and Eric Horvitz. Beyond the Commons: Investigating the Value of Personalizing Web Search. Workshop on New Technologies for Personalized Information Access (PIA 2005).

[12] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Personalizing Search via Automated Analysis of Interests and Activities. In Proceedings of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 2005. ACM Press.

[13] Jansen, B. J. and Pooch, U. 2000. Web user studies: A review and framework for future work. Journal of the American Society of Information Science and Technology. 52(3), 235 – 246.

[14] Jian-Tao Sun, Hua-Jun Zeng, Huan Liu, Yuchang Lu, and Zheng Chen. CubeSVD: A Novel Approach to Personalized Web Search. WWW 2005, May 10-14, 2005, Chiba, Japan.

[15] Ji-Rong Wen, Jian-Yun and Hong-Jiang Zhang, Query Clustering Using User Logs, ACM Transactions on Information Systems (ACM TOIS), 20(1), 59-81, January, 2002.

[16] MSN Search. http://search.msn.com

[17] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines. Cambridge, University Press, 2000

[18] Open Directory Project. http://dmoz.org/

[19] Rose, D.E. Reconciling Information-Seeking Behavior with Search User Interfaces for the Web. Journal of the American Society of Information Science and Technology

[20] U. Lee, Z. Liu, and J. Cho. Automatic identification of user goals in web search. Technical report, UCLA Computer Science, 2004.

[21] Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W., and Li, Y. 2005. Detecting dominant locations from search queries. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05. ACM Press, New York, NY, 424-431.

[22] Yahoo Directory. http://search.yahoo.com/dir