Detecting Multilingual and Multi-Regional Query Intent in Web Search

Yi Chang, Ruiqiang Zhang, Srihari Reddy
Yahoo! Labs
701 First Avenue
Sunnyvale, CA 94089
{yichang,ruiqiang,sriharir}@yahoo-inc.com

Yan Liu
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
yanliu.cs@usc.edu

Abstract

With rapid growth of commercial search engines, detecting multilingual and multi-regional intent underlying search queries becomes a critical challenge to serve international users with diverse language and region requirements. We introduce a query intent probabilistic model, whose input is the number of clicks on documents from different regions and in different language, while the output of this model is a smoothed probabilistic distribution of multilingual and multi-regional query intent. Based on an editorial test to evaluate the accuracy of the intent classifier, our probabilistic model could improve the accuracy of multilingual intent detection for 15%, and improve multi-regional intent detection for 18%. To improve web search quality, we propose a set of new ranking features to combine multilingual and multi-regional query intent with document language/region attributes, and apply different approaches in integrating intent information to directly affect ranking. The experiments show that the novel features could provide 2.31% NDCG@1 improvement and 1.81% NDCG@5 improvement.

Introduction

Ranking is the core technology for a commercial search engine, and the framework of learning to rank (Liu 2009) has been widely applied in many commercial search engines. Given a query, a ranking function measures the relevance of each document to the query, sorts all documents based on their relevance scores, and then presents a list of top-ranked ones to the user. Accuracy of ranking function is heavily dependent on training data, ranking features, and ranking algorithms.

Understanding the intent underlying user queries could help customizing search results and improve user satisfaction. In this paper, detecting multi-regional intent refers to predicting the possible regions of expected URLs, here we restrict regions as countries or territories; correspondingly, detecting multilingual intent refers to predicting possible language of the expected results. As a commercial search engine needs to serve international users with diverse region and language requirements, understanding multilingual and multi-regional intent underlying search queries becomes a critical challenge, while fail to present URLs in expected languages or from expected regions would severely damage the search user experience. For example, a query *google* could point to different landing pages if it is issued in different region: a US user would expect http://www.google.com; a user in Japan would want http://www.google.co.jp in Japanese; while an Indian user would prefer http://www.google.com instead of http://www.google.co.in, since click log shows 42% more clicks on the former URL than the latter by Indian users. Things are more complex as different language share with same script but with ambiguous meaning. For example, given a query *LV*, American users could refer to either *Las Vegas* or *Louis Vuitton*, while majority of Chinese users only refer to the brand of *Louis Vuitton*. In the other word, multilingual and multi-regional query intent is not only related to the query, but also related to the region which search users are located in.

Previous works demonstrate that click-through data could provide rich information to improve relevance of search results (Radlinski and Joachims 2007), (Joachims 2002). We also try to leverage click-through data to estimate multilingual intent and multi-regional intent of each query, since the regions and languages of clicked documents would be a reliable source of implicit feedback from users. In this paper, we analyze click-through data to build a probabilistic query intent model, which predicts the preferences of users for those clicked documents with different regions and languages. To predict the intent of unseen queries, a language model based smoothing technique is introduced to increase the coverage of query intent from short and possibly ambiguous query terms.

To improve search relevance, the next challenge is how to optimally customize search results on the basis of multilingual and multi-regional intent detection. In this paper, we explore several techniques to integrate the query language and region intent into ranking functions, which demonstrate significant improvements over a state-of-the-art ranking system. Our major contributions include (i) a hybrid multilingual and multi-regional intent model, which combines both click-through data and language model based smoothing technique; (ii) integrating query intent understanding into ranking features, which significantly improve web search ranking.

As the remaining parts of the paper, we first describe some

related works in next section; then we describe how to build a query intent model from user clicks and smoothing the model with language models, with the evaluation of model accuracy; in the following section, we describe methods of integrating the query intent model outputs to improve ranking, with ranking experimental results. Finally, conclusions and future works are summarized in the last section.

Related Work

Previous work in related areas has been extensively explored on navigational, commercial and Geo intent detection. Jansen *et al.* (Jansen, Booth, and Spink 2008) present query intent classification and study the hierarchy and distribution of intents. Yi *et al.* (Yi, Raghavan, and Leggetter 2009) focus on city level language models for detection and prediction of city level Geo intent of queries. Dai *et al.* (Dai et al. 2006) build models for both web document and query commercial intent detection. Their approach is based on the results on a search page only and does not use clicks. Dong *et al.* (Dong et al. 2010) present ranking approaches to time sensitive and *buzz* queries which uses a query classifier for detecting recency. Ashkan *et al.* (Ashkan et al. 2008) use search ads data to train a decision tree to classify all queries as commercial/noncommercial and natigational/informational.

There exists a couple of related works using clicks to detect region or language intent. Vadrevu *et al.* (Vadrevu et al. 2008) present different features to identify regional sensitivity of a queries based on query co-ocurrence with locations, regional click rate and switching rate of users between regional and global search preferences, and their location likelihood is based on a single language model for n-grams containing location terms. Ceylan *et al.* (Ceylan and Kim 2009) present approaches to build query language classifiers considering the language of the clicked documents, and use linguistic and click features to develop a decision tree classifier for query language identification and intent.

The field of utilizing user clicks for refining and customizing ranking algorithms has been largely focussed on modeling user behavior to obtain query-document level features, or even learning targets for rankings (Dupret and Liao 2010). Li *et al.* (Li, Wang, and Acero 2008) use click graphs to improve coverage of query intent classifiers for vertical search applications, such as, product search and job search. Given the query intent classification result, Bian *et al.* (Bian et al. 2010) propose a learning framework to incorporate query difference to improve ranking function by introducing query dependent loss functions.

Multilingual and Multi-Regional Query Intent Detection

Motivation

Generally speaking, most search users expect the search results would be the URLs that are localized from the user's same location and in the user's same language. That is to say, multilingual and multi-regional query intent detection is not only related to the query, but also related to the region which search users are located in. In the rest of the paper, all experiments and analysis are based on users data from Taiwan, since the task of multilingual and multi-regional query intent detection is most complex and difficult in Taiwan: Taiwan users would accept content from Taiwan, China mainland, Hong Kong, Japan and the US; while Taiwan users could read Traditional Chinese, Simplified Chinese, Japanese and English. However, our approach is general and applicable to users from any other regions as well.

Neither arbitrarily set the multi-regional intent of each query to be the region that the users are from, nor set the multilingual intent to be the official language in that region, since there exist a significant amount of queries, both implicitly and explicitly, targeting documents from other regions or in other languages. For example, the query *2010 expo* from Taiwan should also contain the official homepage of EXPO 2010 from China mainland. Therefore, we formulate the problem of multilingual and multi-regional query intent detection as following: Given a query q, we need to derive the probabilities $P(r_i|q)$ and $P(l_i|q)$, which quantify a user's information need from different regions r_i and in different languages l_i. Technically, we build the query intent detection model in two steps: (i) give a query, we extract number of clicks on documents from different regions and in different language; (ii) we build a language model to handle unseen queries from click-through data, and the same language model is also used to smooth the estimates of existing queries.

Extract Multilingual and Multi-Regional Intent from Click-Through Data

Clicks is one of the most rich resources to obtain user's implicit feedback for a search engine. Besides relevance preference, users also provide the multilingual and multi-regional intent of their queries via clicks, since those clicked documents are most likely to be correlated to user's information need. Therefore, from the attributes of clicked documents, we might find out language or region intent of a given query.

Click-through data contains a list of click events, and each click event can be represented by a tuple (q, url, pos), where q is the query string, url is the URL of clicked document, pos is ranking position of the URL. For a commercial search engine, during crawling or indexing, each URL has been tagged with one or multiple regions according to domain name and IP address, and also been tagged with one or multiple languages according to language identification tools based on page content analysis. We aggregate click event tuple into query click table $(q, c(r_i, q), c(l_j, q))$, where $c(r_i, q)$ is the number of clicks on documents that are tagged with region r_i, and $c(l_j, q)$ is number of clicks on documents that are tagged with language l_j. To obtain statistically significant estimation, queries which have less than 10 clicks are ignored from the query click table, and for those clicked documents with ranking position larger than 10 are also ignored.

From query click table, one can compute the query region

intent probability as:

$$P_{click}(r_i|q) = \frac{c(r_i,q)}{\sum_j c(r_j,q)}, \quad (1)$$

and the query language intent probability as:

$$P_{click}(l_i|q) = \frac{c(l_i,q)}{\sum_j c(l_j|q)}. \quad (2)$$

Obviously, most queries in Taiwan have Taiwan region intent and Traditional Chinese language intent. However, there are a notable proportion of queries that require specific handling due to their special language intent or region intent which are different from the dominant intent of most Taiwan queries. From a click log of 6-month period, we observe only 84.2% of clicked documents from Taiwan region, and 94.6% of clicked documents in Traditional Chinese.

Table 1 lists the multilingual intent and multi-regional intent distributions for several queries extracted from the 6-month click-through data. In this table, *TW* stands for Taiwan, *CN* for China mainland, *HK* for Hong Kong, *JP* for Japan; *ZH_TW* for Traditional Chinese, *ZH_CN* for Simplified Chinese, *JA* for Japanese, *EN* for English; *OTHER* for all the other languages or regions. From Table 1, we notice that query *Hang Seng Index* in Tradition Chinese has superior Hong Kong region intent, since Hang Seng is the Hong Kong stock market index. This query also has dominant Traditional Chinese intent because Traditional Chinese is the most popular language in Hong Kong. Although the query *Beijing University* is written in Traditional Chinese, most of the user still prefer those documents written in Simplified Chinese, since Beijing University is located in China mainland which uses Simplified Chinese. Similar pattern can be applied to queries, such as *2008 Olympics* and *CNN*.

Query Intent Language Model

Extracting multilingual and multi-regional query intent from click-through data can yield strong signals for popular queries. These queries have large number of clicks which indicate query intent with high confidence. For non-popular queries or for unseen queries which users did not issue before, click-through data does not contain query intent information. In these cases, we use language models to infer intent distributions. This is done by extrapolating query intent from clicks for queries observed in the click-through data.

We built n-gram language models (LM) for multilingual and multi-regional query intent classification using click probabilities. Our work is similar to (Ceylan and Kim 2009), yet the data we used to train LMs are from user's clicks. Tuples $(q, c(r_i, q))$ and $(q, c(l_j, q))$ denote the number of clicks on documents from i-th region and in j-th language, respectively. We first decompose the query q into all possible n-grams, and each n-gram for the query is assumed to have the same click counts as the original query. For example, if query q has two words 't_1t_2', in a word-based LM the decomposition will replace the original row of query 't_1 t_2' with three rows: 't_1', 't_2', 't_1t_2', and each row has the same click counts $c(r_i, q)$ and $c(l_j, q)$. Then we aggregate the table by collapsing identical n-grams into a single row by adding their click counts together. The aggregated n-gram table has the form of $(t, c_{lm}(r_i, t), c_{lm}(l_j, t))$, where t denotes one of the n-grams, $c_{lm}(r_i, t)$ means the number of clicks on the documents from region r_i for the query containing t, and $c_{lm}(l_j, t)$ refers to the number of clicks on the documents in language l_j for the query containing t.

The n-gram table thus built is used to train an intent LM. The basic formulation is:

$$P(t_m|t_{m-1}\cdots t_{m-n+1}, l_j) = \frac{c_{lm}(l_j, t_m\cdots t_{m-n+1})}{c_{lm}(l_j, t_{m-1}\cdots t_{m-n+1})}. \quad (3)$$

This basic formulation needs to be smoothed to compensate unseen n-grams, with Good-Turing method (Katz 1987). If there is no clicks for n-gram $t_{m-1}\cdots t_{m-n+1}$, the calculation of probability $P(t_m|t_{m-1}\cdots t_{m-n+1}, l_j)$ is back-off to $P(t_m|t_{m-1}\cdots t_{m-n+2}, l_j)$.

There are several differences between our query intent LM and commonly used LM, which is concerned with assigning probabilities to future words conditioned on words seen so far. First of all, in classical LM the data for building model is a tokenized text collection or corpus, while for the query intent model the corpus only consists of queries in the click log. Secondly, the n-gram frequencies for a classical language model are the counts of n-grams in the corpus, but for the query intent model, we use number of clicks for frequencies of query n-grams. The reliance on clicks adds confidence to our intent modeling. For example, even for a query with high frequency, if the current search result does not detect the intent, it may get reflected in the click counts. A query for which the clicks are distributed on documents of different regions would indicate no strong regional intent. This matches with the intuition: click counts for a specific language or region for a given query indicates user's preference for that language or region.

Since in many regions, such as China mainland, Taiwan, Hong Kong and Japan, query logs consist of both English queries and queries in Asian scripts, we build word-based LMs depending on the unit of segmentation. For each language-dependent n-gram table extracted from click-through data, we have an query intent estimate based on an LM. From this LM, we use Bayes rule to find the language intent of query q, i.e.,

$$P_{LM}(l_i|q) \propto P(q|l_i)P(l_i), \quad (4)$$

and the standard chain rule is applied to $P(q|l_i)$. For example, if the query q consists of term sequence $t_0t_1\cdots t_m$, then

$$P(q|l_i) = P(t_0|l_i)P(t_1|t_0, l_i)\cdots P(t_m|t_{m-1}\cdots t_{m-n+1}, l_i), \quad (5)$$

where $P(t_m|t_{m-1}\cdots t_{m-n+1}, l_i)$ is the n-gram LM for language intent l_i, and is evaluated via (3). For our word-based LMs, we choose $n = 3$. In addition, $P(l_i)$ is the prior probability for a given language intent l_i. It can be estimated as the proportion of aggregated number of clicks on documents of a given language normalized by clicks on all documents. The same method was also applied for region intent $P_{LM}(r_j|q)$.

Query	Multi-Regional Intent						Multilingual Intent				
	TW	CN	HK	JP	US	OTHER	ZH_TW	ZH_CN	JA	EN	OTHER
Hang Seng Index (in ZH_TW)	0.14	0.04	0.68	0	0.14	0	0.86	0	0	0	0.14
Beijing University (in ZH_TW)	0.1	0.88	0	0	0.02	0	0.09	0.88	0	0	0.03
CNN	0.0	0.03	0.0	0	0.97	0	0.03	0.0	0.0	0.97	0.0
2008 Olympics	0	0.82	0	0	0.18	0	0	0.43	0	0.57	0

Table 1: Examples of multilingual and multi-regional query intent extracted from Taiwan click-through data.

Our final multilingual and multi-regional query intent model consists of two factors. For high frequency queries with sufficient click information, intent is derived from the past user behaviors. For those queries where the confidence from the click signals is not strong, we use a weighted combination of historical preference from click through data for the exact query and a smoothed preference from language models. The intent from the language model is not only based on the issued query, but also every n-gram in the query as described earlier. Finally, for unseen queries we totally rely on the language model. Succinctly,

$$P(l_i|q) = P_{click}(l_i|q) + \frac{\lambda}{(1+log(1+freq(q))} P_{LM}(l_i|q) = \frac{c(l_i,q)}{\sum_j c(l_j|q)} + \frac{\lambda}{(1+log(1+freq(q))} P_{LM}(l_i|q) \quad (6)$$

Similarly,

$$P(r_i|q) = \frac{c(r_i,q)}{\sum_j c(r_j|q)} + \frac{\lambda}{(1+log(1+freq(q))} P_{LM}(r_i|q) \quad (7)$$

where $freq(q)$ is the frequency of the query in the query log and λ is tuned on a labeled hold out set.

Query Intent Model Evaluation

To evaluate our multilingual and multi-regional query intent model, we build a golden set of human labeled data. We randomly selected about 6,000 queries, and request editors to provide the ground truth of multilingual and multi-regional intent for each query.

The complexity of the problem can be gauged by a naive intent model which always predicts Taiwan for regional intent and Traditional Chinese for language intent since these are the dominant query intent of Taiwan users. By comparing the predictions of this naive model with the editorial labels, we found that this naive model has an accuracy of 62% for multi-regional intent and 69% for multilingual intent, which can be regarded as the baseline.

The naive model does not depend on any click information. Although our full multilingual and multi-regional query intent model, as (6) and (7), consists of two components: one is purely from click statistics and the other is based on query term smoothing, it is worthwhile to see how far we can get purely based on the click statistics, such as $P_{click}(l_i|q)$ and $P_{click}(r_i|q)$, and how much the full query intent models can achieve due to query smoothing. To evaluate the accuracies of the different models discussed above, we rank the intents estimated from differen methods and compare the predicted top intent against the top intent tagged by the editors, and the accuracy are averaged over all 6,000 queries.

Table 2 reports the accuracy of different approaches for query intent classification. As we can see, generally speaking, the full LM intent models outperform the baseline on all metrics: for multilingual intent prediction, LM based model could achieved 84% accuracy, which is 15% improvement over the baseline; while for multi-regional intent prediction, LM based model could achieved 80% accuracy, which is 18% improvement over the baseline. Comparing with click-based approach, the full LM intent models are 4% better than the purely click-based approach for multilingual intent prediction, but on par for multi-regional intent prediction.

Incorporating Multilingual and Multi-Regional Intent for Ranking

Converting Query Intent Features

Most of existing commercial search engine already utilized the region and language information for each URL as different ranking features. Converting multilingual and multi-regional query intent into ranking features could align the intent information from both query and document perspective.

We design a set of features that are obtained from the multilingual and multi-regional intent model. These features are listed as below:

i. Query multi-regional intent probability: $[q_{r_1}, q_{r_2}, \cdots, q_{r_n}]$
where n is the number of popular regions of intent for users in a region, i.e., $q_{r_i} = P(r_i|q)$.

ii. Query multilingual intent probability: $[q_{l_1}, q_{l_2}, \cdots, q_{l_m}]$
where m is the number of popular languages of intent for users in a region, i.e., $q_{l_i} = P(l_i|q)$.

iii. Document regions: $[d_{r_1}, d_{r_2}, \cdots, d_{r_N}]$
where N is the set of all regions identified by the search engine at index time. Each d_{r_i} is binary valued. A document can have non-zero document region feature value for more than one region.

iv. Document languages: $[d_{l_1}, d_{l_2}, \cdots, d_{l_M}]$
where M is the set of all languages identified by the search engine at index time. Each d_{l_i} is binary valued with a document having a non-zero document language feature value for only one language.

v. Query-Document multi-regional similarity: qd_{rsim}
This is a simple measure of similarity between the document region vector and the query multi-regional intent

Method	Multi-Regional Intent						Multilingual Intent				
	TW	CN	HK	JP	US	Overall	ZH_TW	ZH_CN	JA	EN	Overall
Naive	0.62	-	-	-	-	0.62	0.69	-	-	-	0.69
Click	0.85	0.76	0.28	0.64	0.65	**0.80**	0.85	0.70	0.69	0.56	**0.80**
Click + LM	0.86	0.74	0.17	0.61	0.60	**0.80**	0.95	0.46	0.27	0.36	**0.84**

Table 2: Accuracy of multilingual and multi-regional intent detection with different methods.

vector. It is defined as the inner product of $[q_{r_1}, q_{r_2}, \cdots, q_{r_n}]$ and $[d_{r_1}, d_{r_2}, \cdots, d_{r_n}]$.

vi. Query-Document multilingual similarity: qd_{lsim}
This is a simple measure of similarity between the document language vector and the query multilingual intent vector. It is defined as the inner product of $[q_{l_1}, q_{l_2}, \cdots, q_{l_m}]$ and $[d_{l_1}, d_{l_2}, \cdots, d_{l_m}]$.

vii. Query-Document multilingual and multi-regional similarity: qd_{rlsim}
This is just the sum of the query-document multi-regional similarity qd_{rsim} and query-document multilingual similarity qd_{lsim}.

Re-rank with Score Adjustment

To improve search ranking with multilingual and multi-regional intent features, one straightforward solution is to adjust the relevance score of each URL calculated by the ranking function as a post processing step. This approach is simple and effective, without requiring any market specific training data.

The score adjustment re-rank approach is based on query-document multi-regional similarity qd_{rsim} and query-document multilingual similarity qd_{lsim}. The final relevance score is designed as:

$$\begin{aligned} \text{Re-rank Score} &= machine\_learned\_score \\ &+ [\alpha_{r_1}, \alpha_{r_2}, ...] * [qd_{rsim_1}, qd_{rsim_2}, ...]^T \\ &+ [\beta_{l_1}, \beta_{l_2}, ...] * [qd_{lsim_1}, qd_{lsim_2}, ...]^T, \end{aligned}$$

where α_{r_i} and β_{l_j} are the weights for document from region r_i and in language l_j respectively, and these parameters can be tuned over the hold out set using limited human labeled data.

Learning to Rank with Query Intent Features

Learning to rank (LTR) represents a class of supervised machine learning tasks with the goal of automatically constructing ranking functions from training data. In our experiments, ranking functions are trained with an algorithm called Gradient Boosting Decision Trees (Friedman 2001), which is proved to be robust and effective in Yahoo! Learning to Rank Challenge in 2010 (Chapelle and Chang 2011).

The baseline is a learning to rank model trained with more than 500 production features from a commercial web search engine. These features can be divided into 9 categories: web map features, document statistics features, document classifier features, query features, text match features, topical match features, click features, external reference features and time sensitive features. The detailed feature explanation about each type of features could be found in (Chapelle and Chang 2011).

To demonstrate the effectiveness of the new features, we build another learning to rank model which utilizes existing baseline features, plus the novel multilingual and multi-regional intent features proposed above.

Our ranking experiments are based on a large collection of human labeled data. The training data consists of 262,512 pairs of query-URL with 8600 unique queries, and each query-URL pair is editorially evaluated for relevance on a scale of 1 to 5.

Ranking Experiments

The final evaluation test to measure NDCG (Jarvelin and Kekalainen 2002) is performed on a testing set, which is consisted of 4140 randomly sampled queries from 1-month period of query log Taiwan users. The total number of labeled query-URL pairs for testing was 101,279.

Table 3 reports NDCG@1 and NDCG@5 on all the 4140 random queries, and Table 4 illustrates NDCG@1 and NDCG@5 on the affected queries, which refer to those queries whose ranking are changed. It is obvious that re-rank based score adjustment provide marginal relevance gain, while leveraging all new features to train a machine learning ranking model could significantly improve NDCG@5 by 1.81%, NDCG@1 by 2.31% on all queries.

To consider the feature importance according to (Hastie, Tibshirani, and Friedman 2001), the new proposed query-document intent features are ranked very high: qd_{rsim} is ranked at position 5, qd_{rlsim} at position 10, qd_{lsim} at position 13, with 57%, 41% and 34% the power of the most important feature, respectively, while query intent features and document intent features are ranked much lower. Thus the feature importance of the LTR model matches very well with our intuition of the need to combine query intent with document intent.

Conclusions and Future Work

In this paper, we present a multilingual and multi-regional query intent model and its application on web search ranking. Our approach combines clicks for popular queries with language models for smoothing unseen queries. We also explore different approaches to incorporate the query intent information into ranking for relevance improvement. According to editorial based experiments, our query intent model could reach more than 80% accuracy, which significantly improves 18% accuracy for multi-regional detection and 15% for multilingual intent detection, comparing with the baseline approach. With regards to applying query intent for ranking, our finding is that a unified learning to rank

Model	NDCG@1	NDCG@5	Percentage Gain	
			NDCG@1	NDCG@5
Baseline	0.735	0.730	-	-
Score Adjustment Re-Rank	0.736	0.731	0.16%	0.14%
LTR with new features	0.752	0.743	**2.31%**	**1.81%**

Table 3: NDCG@1 and NDCG@5 for different approaches on all 4140 random queries. Bold font means NDCG gain are statistically significant with p-value smaller than 0.01

Model	NDCG@1	NDCG@5	Percentage Gain		Affected Queries & Coverage
			NDCG@1	NDCG@5	
Score Adjustment Re-Rank	0.682	0.670	0.43%	1.48%	444 (10.7%)
LTR with new features	0.724	0.710	**3.30%**	**2.72%**	2708 (65.4%)

Table 4: NDCG@1 and NDCG@5 for different approaches on affected queries only. Bold font means NDCG gain are statistically significant with p-value smaller than 0.01

approach of using existing features plus new intent features could significantly outperform external score adjustment re-rank method.

Although our intent model manages to capture intent from user's click data, it ignores session level click information, query reformulation and document positions. Clearly, these information can be very useful for intent modeling. We plan to improve our multilingual and multi-regional intent model by considering these issues in the future. Another area that has not been well explored in large scale settings is on-line learning for intent detection for web search ranking. There is also strong evidence of certain vertical intents also having strong correlation with region and language intents, e.g. commercial intent is invariably also has a region intent. We plan to investigate these issues in our future work.

References

Ashkan, A.; Clarke, C.; Agichtein, E.; and Guo, Q. 2008. Characterizing query intent from sponsored search clickthrough data. In *Proceedings of the SIGIR Workshop on Information Retrieval in Advertising*.

Bian, J.; Liu, T.-Y.; Qin, T.; and Zha, H. 2010. Ranking with query-dependent loss for web search. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*.

Ceylan, H., and Kim, Y. 2009. Language identification of search engine queries. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 1066–1074.

Chapelle, O., and Chang, Y. 2011. Yahoo! learning to rank challenge overview. *Journal of Machine Learning Research* Proceedings Track(14):1–24.

Dai, H.; Zhao, L.; Nie, Z.; Wen, J.-R.; Wang, L.; and Li., Y. 2006. Detecting online commercial intention. In *Proceedings of the 15th international conference on World wide web*, 829¨C837.

Dong, A.; Chang, Y.; Zheng, Z.; Mishne, G.; Bai, J.; Zhang, R.; Buchner, K.; Liao, C.; and Diaz, F. 2010. Towards recency ranking in web search. In *WSDM*, 11–20.

Dupret, G., and Liao, C. 2010. Cumulated relevance: A model to estimate document relevance from the clickthrough logs of a web search engine. In *To appear in Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM)*.

Friedman, J. 2001. Greedy function approximation: a gradient boosting machine. *Ann. Statist.* 29:1189–1232.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Jansen, B. J.; Booth, D.; and Spink, A. 2008. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.* 44(3):1251¨C1266.

Jarvelin, K., and Kekalainen, J. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems* 20:422–446.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Katz, S. M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing* 35(3):400–401.

Li, X.; Wang, Y.-Y.; and Acero, A. 2008. Learning query intent from regularized click graphs. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 339–346.

Liu, T. Y. 2009. *Learning to Rank for Information Retrieval*. Foundation and Trends on Information Retrieval.

Radlinski, F., and Joachims, T. 2007. Active exploration for learning rankings from clickthrough data. *Proc. of ACM SIGKDD Conference*.

Vadrevu, S.; Zhang, Y.; Tseng, B.; Sun, G.; and Li, X. 2008. Identifying regional sensitive queries in web search. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, 1185–1186.

Yi, X.; Raghavan, H.; and Leggetter, C. 2009. Discovering user's specific geo intention in web search. In *Proceedings of the 18th international conference on World wide web*, 481–490.