Query Ambiguity Revisited: Clickthrough Measures for Distinguishing Informational and Ambiguous Queries

Yu Wang
Math & Computer Science Department
Emory University
yu.wang@emory.edu

Eugene Agichtein
Math & Computer Science Department
Emory University
eugene@mathcs.emory.edu

Abstract

Understanding query ambiguity in web search remains an important open problem. In this paper we reexamine query ambiguity by analyzing the result clickthrough data. Previously proposed clickthrough-based metrics of query ambiguity tend to conflate informational and ambiguous queries. To distinguish between these query classes, we introduce novel metrics based on the entropy of the click distributions of individual searchers. Our experiments over a clickthrough log of commercial search engine demonstrate the benefits of our approach for distinguishing informational from truly ambiguous queries.

1 Introduction

Since query interpretation is the first crucial step in the operation of the web search engines, more reliable query intent classification, such as detecting whether a query is ambiguous, could allow a search engine to provide more diverse results, better query suggestions, or otherwise improve user experience.

In this paper we re-examine query ambiguity in connection with searcher clickthrough behavior. That is, we posit that clickthrough information could provide important evidence for classifying query ambiguity. However, we find that previously proposed clickthrough-based measures tend to conflate informational and ambiguous queries. We propose a novel clickthrough measure for query classification, *user click entropy*, and show that it helps distinguish between informational and truly ambiguous queries.

Previous research on this topic focused on binary classification of query ambiguity. Notably, (Teevan et al., 2008) used click entropy as a proxy for query ambiguity to estimate the potential for search personalization. (Mei and Church, 2008) considered click entropy as measure of search difficulty. More broadly, clickthrough information has been used for many other tasks such as improving search ranking (Zhu and Mishne, 2009), learning semantic categories (Komachi et al., 2009), and for topical query classification (Li et al., 2008). However, our work sheds new light on distinguishing between informational and ambiguous queries, by using clickthrough data. Our contributions include:

- More precise definition of query ambiguity in terms of clickthrough behavior (Section 2).
- Entropy-based formalization of resulting click behaviors (Section 3).
- Empirical validation of our methods on a large real query and clickthrough log (Section 4).

2 Defining Query Ambiguity

In this study we focus on two orthogonal query intent dimensions, adapted from the top level of user goal taxonomies such as (Rose and Levinson, 2004). Specifically, a query could be *ambiguous* or *unambiguous*; as well as *informational* or *navigational*. Consider the example queries of each type below:

	Ambiguous	Unambiguous
Informational	"al pacino"	"lyrics"
Navigational	"people"	"google"

The query "al pacino", the name of a famous actor, is a typical ambiguous and informational query. In the clickthrough logs that we examined, the most popular searcher destinations include sites with pictures of Al Pacino, movie sites, and biography sites – corresponding to different informational intents. In contrast, the query "lyrics" has an unambiguous informational intent, which is to explore websites with song lyrics. For the ambiguous navigational query "people", popular destinations include people.com, Yahoo People or People's United Bank. Finally, the

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 361–364,
Los Angeles, California, June 2010.

query "google" is unambiguous and navigational, with over 94% of the clicks on the Google's homepage.

Definitions of query classes: we now more formally define the query classes we consider:

- **Clear**: Unambiguous navigational query, such as "google".
- **Informational**: Unambiguous informational query, such as "lyrics"
- **Ambiguous**: Ambiguous informational or navigational query, such as "people" or "al pacino".

The key challenge in distinguishing the last two classes, Informational and Ambiguous, is that the overall clickthrough patterns for these classes are similar: in both cases, there are clicks on many results, without a single dominant result for the query.

3 Clickthrough Measures for Distinguishing Ambiguous and Informational Queries

In this section we describe the features used to represent a query for intent classification, listed in Table 1. In addition to popular features such as clickthrough frequency and query length, we introduce novel features related to user click entropy, to capture the distinction between informational and ambiguous queries.

Overall Entropy: Previous methods for query classification utilize entropy of all result clicks for a query, or overall entropy (the uncertainty associated with obtaining a click on any specific result), defined as:

$$H(R_q) = -\sum_{r \in R_q} p(r) \log p(r)$$

R_q is the set of results r, clicked by all users after submitting the query q. For example, a clear query "target" has the overall entropy of 0.36, and most results corresponding to this query point to Target's company website. The click log data shows that 85% of the users click the Target website for this query. In contrast, an unclear query "lyrics" has the overall entropy of 2.26. However, overall entropy is insufficient for distinguishing between informational and ambiguous queries. To fill this gap, we introduce new clickthrough metrics to detect such ambiguous queries.

User Entropy: Recall, that both informational queries and ambiguous queries could have high overall entropy, making it difficult to distinguish them. Thus, we introduce a new metric, *user entropy of a query* q $H(U_q)$, as the average entropy of a distribution of clicks for each *searcher*:

$$H(U_q) = \frac{-\sum_{u \in U_q} \sum_{r \in R_u} p(r) \log p(r)}{|U_q|}$$

where U_q is the set of users who have submitted the query q, and R_u is the set of results r, clicked by the user u. For the example informational query "lyrics", a single user may click many different URLs, thereby increasing user entropy of this query to 0.317. While for an ambiguous query, which has multiple meanings, a user typically searches for only one meaning of this query at a time, so the results clicked by each user will concentrate on one topic. For example, the query "people" is ambiguous, and has the overall entropy of 1.73 due to the variety of URLs clicked. However, a particular user usually clicks only one of the websites, resulting in low *user entropy* of 0.007. Figure 1 illustrates the difference in the distributions of informational and ambiguous queries according to their overall and user entropy values: more informational queries tend to have medium to high User Entropy values, compared to the truly ambiguous queries.

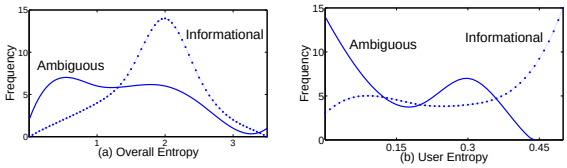

Figure 1: Frequency of ambiguous and informational queries by Overall Entropy (a) and User Entropy (b).

Domain Entropy: One problem with the above measures is that clickthrough data for individual URLs is sparse. A common approach is to *backoff* to the URLs domain, with the assumption that URLs within the same domain usually relate to the same topic or concept. Therefore, domain entropy $H(D_q)$ of a query may be more robust, and is defined as:

$$H(D_q) = -\sum_{d \in D_q} p(d) \log p(d)$$

where D_q are the domains of all URL clicked for q. For example, the query "excite" is a navigational and clear query, as all the different clicked URLs for this query are within the same domain, *excite.com*.

Query Feature	Description
QueryLength	Number of tokens (words) in the query
ClickFrequency	Number of total clicks for this query
OverallEntropy	Entropy of all URLs for this query
UserEntropy*	Average entropy of the URLs clicked by one user for this query
OverallDomainEntropy	Entropy of all URL domains for this query
UserDomainEntropy*	Average entropy of URL domains clicked by one user for this query
RelativeUserEntropy*	Fraction of UserEntropy divided by OverallEntropy
RelativeOverallEntropy*	Fraction of OverallEntropy divided by UserEntropy
RelativeUserDomainEntropy*	Fraction of UserDomainEntropy divided by OverallDomainEntropy
RelativeOverallDomainEntropy*	Fraction of OverallDomainEntropy divided by UserDomainEntropy

Table 1: Features used to represent a query (* indicates features derived from User Entropy).

While this query has high Overall and User Entropy values, the Domain Entropy is low, as all the clicked URLs for this query are within the same domain.

The features described here can then be used as input to many available classifiers. In particular, we use the Weka toolkit[1], as described below.

4 Experimental Results

We describe the dataset and annotation process, and then present and analyze the experimental results.

Dataset: We use an MSN Search query log (from 2006 Microsoft data release) with 15 million queries, from US users, sampled over one month. Queries with click frequency under 10 are discarded. As a result, 84,703 unique queries remained, which form our universe of queries. To separately analyze queries with different frequencies, we divide the queries into three groups: low frequency group (10-100 clicks), medium frequency group (100-1000 clicks) and high frequency group (over 1000 clicks). From each group, we draw a random sample of 50 queries for manual labeling, for the total of 150 queries. Each query was labeled by three members of our lab. The inter-annotator agreeement was 85%, and Cohen's Kappa value was 0.77.

Table 2 reports the distribution of query classes in our dataset. Note that low frequency queries dominate, but are equally represented in the data samples used for classification training and prediction (we will separately analyze performance on different query frequency groups).

Results: Table 3 shows that best classification required User Entropy features. The Weka classifiers were Naive Bayes (NB), Logistic Regression (Logistic), and Support Vector Machines (SVM).

[1]http://www.cs.waikato.ac.nz/ml/weka/

	Clear	Informational	Ambiguous	Frequency (%)
High	76%	8%	16%	255 (0.3%)
Medium	52%	20%	28%	3802 (4.5%)
Low	32%	46%	22%	80646 (95.2%)

Table 2: Frequency distribution of different query types

	All	Clear		Informational		Ambiguous	
	Ac.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.
All features							
NB	0.72	**0.90**	0.85	**0.77**	0.54	0.42	**0.61**
Logistic	**0.77**	0.84	0.98	0.68	0.73	0.59	0.30
SVM	0.76	0.79	1.00	0.69	0.78	**0.71**	0.15
Without user entropy							
NB	0.73	0.85	0.95	0.63	0.73	0.39	0.21
Logistic	0.73	0.84	0.95	0.63	0.68	0.47	0.27
SVM	0.74	0.79	1.00	0.65	0.76	0.50	0.09

Table 3: Classification performance by query type

	High	Mid	Low		
	Ac.	Ac.	Ac.	Pre.	Rec.
All features					
NB	0.76	0.76	0.74	**0.80**	**0.74**
Logistic	0.78	0.76	0.70	0.68	0.7
SVM	0.78	0.72	**0.79**	0.69	0.72
Without user entropy					
NB	0.80	0.76	0.70	0.66	0.70
Logistic	0.80	**0.82**	0.66	0.63	0.66
SVM	0.80	0.78	0.68	0.62	0.68

Table 4: Classification performance by query frequency

Recall, that low frequency queries dominate our dataset, so we focus on performance of low frequency queries, as reported in Table 4. The respective χ^2 values are reported in (Table 5). The features *UserDomainEntropy* and *UserEntropy* correlate the most with manual query intent labels.

As an alternative to direct multiclass classification described above, we first classify clear vs. unclear queries, and only then attempt to distinguish ambiguous and informational queries (within the un-

Feature	χ^2 (*multiclass*)	χ^2 (*binary*)
UserDomainEntropy	132.9618	23.3629
UserEntropy	128.0111	21.6112
RelativeOverallEntropy	96.6842	20.0255
RelativeUserEntropy	98.6842	20.0255
OverallEntropy	96.1205	0

Table 5: χ^2 values of top five features for *multiclass* classification (*clear* vs. *informational* vs. *ambiguous*) and for and for *binary* classification (*informational* vs. *ambiguous*), given the manual *unclear* label.

	Overall	*Informational*		*Ambiguous*	
	Ac.	Pre.	Rec.	Pre.	Rec.
With User Entropy features					
NB	**0.72**	**0.82**	0.60	0.65	**0.85**
Logistic	0.71	0.74	0.70	0.69	0.73
SVM	0.65	0.64	0.73	0.64	0.55
Without User Entropy features					
NB	0.66	0.65	0.76	0.67	0.55
Logistic	0.68	0.69	0.73	0.68	0.64
SVM	0.68	0.67	**0.81**	**0.72**	0.55

Table 6: Binary classification performance for queries manually labeled as unclear.

clear category). For classification between clear and unclear queries, the accuracy was 90%, precision was 91%, and recall was 90%. The results for subsequently classifying ambiguous vs. information queries are reported in Table 6. For this task, User Entropy features are beneficial, while the χ^2 value or Overall Entropy is 0, supporting our claim that User Entropy is more useful for distinguishing informational from ambiguous queries.

Discussion: Interestingly, User Entropy does not show a large effect on classification of High and Medium frequency queries. However, as Table 2 indicates, High and Medium frequency queries are largely *clear* (76% and 52%, respectively). As discussed above, User Entropy helps classify unclear queries, but there are fewer such queries among the High frequency group, which also tend to have larger click entropy in general.

An *ambiguous* query is difficult to detect when most users interpret it only one way. For instance, query "ako" was annotated as *ambiguous*, as it could refer to different popular websites, such as the site for Army Knowledge Online and the company site for A.K.O., Inc. However, most users select the result for the Army Knowledge Online site, making the overall entropy low, resulting in prediction as a *clear* query. On the positive side, we find that User Entropy helps detect ambiguous queries, such as "laguna beach", which was labeled *ambiguous* as it could refer to both a geographical location and a popular MTV show. As a result, while the Overall Entropy value of the clickthrough is high, the low User Entropy value identifies the query as truly ambiguous and not informational.

In summary, our techniques are of most help for Low frequency queries and moderately helpful for Medium frequency queries. These results are promising, as Low frequency queries make up the majority of queries processed by search engines, and also contain the highest proportion of informational queries, which our techniques can identify.

5 Conclusions

We explored clickthrough-based metrics for distinguishing between ambiguous and informational queries - which, while exhibiting similar *overall* clickthrough distributions, can be more accurately identified by using our User Entropy-based features. We demonstrated substantial improvements for *low-frequency* queries, which are the most frequent in query logs. Hence, our results are likely to have noticeable impact in a real search setting.

Acknowledgments: This work was partially supported by grants from Yahoo! and Microsoft.

References

M. Komachi, S. Makimoto, K. Uchiumi, and M. Sassano. 2009. Learning semantic categories from clickthrough logs. In *Proc. of ACL-IJCNLP*.

X. Li, Y.Y. Wang, and A.Acero. 2008. Learning query intent from regularized click graphs. In *SIGIR*, pages 339–346.

Q. Mei and K. Church. 2008. Entropy of search logs: how hard is search? with personalization? with backoff? In *Proc. of WSDM*, pages 45–54.

D. E. Rose and D. Levinson. 2004. Understanding user goals in web search. In *Proc. of WWW*, pages 13–19.

J. Teevan, S. T. Dumais, and D. J. Liebling. 2008. To personalize or not to personalize: modeling queries with variation in user intent. In *Proc. of SIGIR*, pages 163–170.

G. Zhu and G. Mishne. 2009. Mining rich session context to improve web search. In *Proc. of KDD*, pages 1037–1046.