Using Word-Sense Disambiguation Methods to Classify Web Queries by Intent

Emily Pitler
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
epitler@seas.upenn.edu

Ken Church
Johns Hopkins University
Human Language Technology Center of Excellence
Baltimore, MD 21211
Kenneth.Church@jhu.edu

Abstract

Three methods are proposed to classify queries by intent (CQI), e.g., navigational, informational, commercial, etc. Following mixed-initiative dialog systems, search engines should distinguish navigational queries where the user is taking the initiative from other queries where there are more opportunities for system initiatives (e.g., suggestions, ads). The query intent problem has a number of useful applications for search engines, affecting how many (if any) advertisements to display, which results to return, and how to arrange the results page. Click logs are used as a substitute for annotation. Clicks on ads are evidence for commercial intent; other types of clicks are evidence for other intents. We start with a simple Naïve Bayes baseline that works well when there is plenty of training data. When training data is less plentiful, we back off to nearby URLs in a click graph, using a method similar to Word-Sense Disambiguation. Thus, we can infer that *designer trench* is commercial because it is close to *www.saksfifthavenue.com*, which is known to be commercial. The baseline method was designed for precision and the backoff method was designed for recall. Both methods are fast and do not require crawling webpages. We recommend a third method, a hybrid of the two, that does no harm when there is plenty of training data, and generalizes better when there isn't, as a strong baseline for the CQI task.

1 Classify Queries By Intent (CQI)

Determining query intent is an important problem for today's search engines. Queries are short (consisting of 2.2 terms on average (Beitzel et al., 2004)) and contain ambiguous terms. Search engines need to derive what users want from this limited source of information. Users may be searching for a specific page, browsing for information, or trying to buy something. Guessing the correct intent is important for returning relevant items. Someone searching for *designer trench* is likely to be interested in results or ads for trench coats, while someone searching for *world war I trench* might be irritated by irrelevant clothing advertisements.

Broder (2002) and Rose and Levinson (2004) categorized queries into those with *navigational*, *informational*, and *transactional* or *resource-seeking* intent. Navigational queries are queries for which a user has a particular web page in mind that they are trying to navigate to, such as *greyhound bus*. Informational queries are those like *San Francisco*, in which the user is trying to gather information about a topic. Transactional queries are those like *digital camera* or *download adobe reader*, where the user is seeking to make a transaction or access an online resource.

Knowing the intent of a query greatly affects the type of results that are relevant. For many queries, Wikipedia articles are returned on the first page of results. For informational queries, this is usually appropriate, as a Wikipedia article contains summaries of topics and links to explore further. However, for navigational or transactional queries, Wikipedia is not as appropriate. A user looking for the *greyhound bus* homepage is probably not interested in facts about the company. Similarly, someone looking to *download adobe reader* will not be interested in Wikipedia's description of the product's history. Conversely, for informational queries, Wikipedia articles tend to be appropriate while advertisements are not. The user searching for *world war I trench* might find the Wikipedia article on trench warfare useful, while he is prob-

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 1428–1436,
Singapore, 6-7 August 2009.

(a) The advertisements and related searches are probably more likely to be clicked on than the top result for *designer trench*.

(b) The top result will receive more clicks than the spelling suggestion. Wikipedia often receives lots of clicks, but not for commercial queries like *bestbuy*.

Figure 1: Results pages from two major search engines. A search results page has limited real estate that must be divided between search results, spelling suggestions, query suggestions, and ads.

ably not interested in purchasing clothing, or even World War I related products. We noticed empirically that queries in the logs tend to have a high proportion of clicks on the Wikipedia article or the ads, but almost never both. The Wikipedia page for Best Buy in Figure 1(b) is probably a waste of space. Knowing whether a particular query is navigational, informational, or transactional would improve search and advertising relevance.

After a query is issued, search engines return a list of results, and possibly also advertisements, suggestions of related searches, and spelling suggestions. For different queries, these alternatives have varying utilities to the users. Consider the queries in Figures 1(a) and 1(b). For *designer trench*, the advertisements may well be more useful to the user than the standard set of results. The query suggestions for *designer trench* all would help refine the query, whereas the suggestions for *bestbuy* are less useful, as they would either return the same set of results or take the user to Best Buy's competitors' sites. The spelling suggestion for *best buy* instead of *bestbuy* is also unnecessary. Devoting more page space to the content that is likely to be clicked on could help improve the user experience.

In this paper we consider the task of: given a class of queries, which types of answer (standard search, ads, query suggestions, or spelling sug-

gestions) are likely to be clicked on? Typos will tend to have more clicks on the spelling suggestions, informational queries will have more clicks on Wikipedia pages, and commercial queries will have more clicks on the ads. The observed behavior of where users click tells us something about the hidden intentions of the users when they issue that query.

We focus on *commercial intent* (Dai et al., 2006), the intent to purchase a product or service, to illustrate our method of predicting query intent. The business model of web search today is heavily dependent on advertising. Advertisers bid on queries, and then the search results page also contains "sponsored" sites by the advertisers who won the auction for that query. It is thus advantageous for the advertisers to bid on queries which are most likely to result in a commercial transaction. If a query is classified as likely implying commercial intent, but the advertisers have overlooked this query, then the search engine may want to suggest that advertisers bid on that query. The search engine may also want to treat queries classified as having commercial intent differently, by rearranging the appearance of the page, or by showing more or fewer advertisements.

This paper starts with a simple Naïve Bayes baseline to classify queries by intent (CQI). Supervised methods work well, especially when there is plenty of annotated data for testing and training. Unfortunately, since we don't have as much annotated data as we might like, we propose two workarounds:

1. Use click logs as a substitute for annotated data. Clicks on ads are evidence for commercial intent; other types of clicks are evidence for other intents.

2. We propose a method similar to Yarowsky (1995) to generalize beyond the training set.

2 Related Work

Click logs have been used for a variety of tasks involved in information retrieval, including predicting which pages are the best results for queries (Piwowarski and Zaragoza, 2007; Joachims, 2002; Xue et al., 2004), choosing relevant advertisements (Chakrabarti et al., 2008), suggesting related queries (Beeferman and Berger, 2000), and personalizing results (Tan et al., 2006). Queries that have a navigational intent tended to have a highly skewed click distribution, while users clicked on a wider range of results after issuing informational queries. Lee et al. (2005) used the click distributions to classify navigational versus informational intents.

While navigational, informational, and resource-seeking are very broad intentions, other researchers have looked at personalization and intent on a per user basis. Downey et al. (2008) use the last URL visited in a session or the last search engine result visited as a proxy for the user's information goal, and then looked at the correspondence between information needs and queries (how the goals are expressed).

We are interested in a granularity of intent in between navigational/informational/resource-seeking and personalized intents. For these sorts of intents, the web pages associated with queries provide useful information. To classify queries into an ontology of commercial queries, Broder et al. (2007) found that a classifier that used the text of the top result pages performed much better than a classifier that used only the query string. While the results are quite good on their hierarchy of 6000 types of commercial intents, they manually constructed about 150 hand-picked examples each for each of the 6000 intents. Beitzel et al. (2005) do semi-supervised learning over the query logs to classify queries into topics, but also train with hundreds of thousands of manually annotated queries. Thus, while we also use the query logs and the identities of web pages of associated with each query, we are interested in finding methods that can be applied when that much annotation is prohibitive.

Semi-supervised methods over the click graph make it possible to train classifiers after starting from a much smaller set of seed queries. Li et al. (2008) used the semi-supervised learning method described in Zhou et al. (2004) to gain a much larger training set of examples, and then trained classifiers for product search or job search on the expanded set. Random walk methods over the click graph have also been used to propagate relations between URLs, for tasks such as finding "adult" content (Craswell and Szummer, 2007) and suggesting related queries (Antonellis et al., 2008) and content (Baluja et al., 2008). In our work we also seek to classify query intent using the click graph, but we demonstrate the effectiveness of a simple method by building deci-

sion lists of URLs. In addition, we evaluate our method automatically by using user click rates, rather than assembling hand-labeled examples for training and testing.

Dai et al. (2006) also classified queries by commercial intent, but their method involved crawling the top landing pages for each query, which can be quite time-consuming. In this paper we investigate the commercial intent problem when crawling pages is not feasible, and use only the identities of the top URLs.

3 Using Click Logs as a Substitute for Annotation

Prior work has used click logs in lieu of manual annotations of relevance ratings, either of webpages (Joachims, 2002) or of sponsored search advertisements (Ciaramita et al., 2008). Here we use the click logs as a large-scale source of intents. Logs from Microsoft's Live Search are used for training and test purposes. Logs from May 2008 were used for training, and logs from June 2008 were used for testing.

The logs distinguish four types of clicks: (a) search results, (b) ads, (c) spelling suggestions and (d) query suggestions. Some prototypical queries of each type are shown in Table 1. As mentioned above, clicks on ads are evidence for commercial intent; other types of clicks are evidence for other intents. The query, *ebay official*, is assumed to be commercial intent, because a large fraction of the clicks are on ads. In contrast, typos tend to have relatively more clicks on "did-you-mean" spelling suggestions.

The query logs contain over a terabyte of data for each day, and our experiments were done using months of logs at a time. We used SCOPE (Chaiken et al., 2008), a scripting programming language designed for doing Map-Reduce (Dean and Ghemawat, 2004) style computations, to distribute the task of aggregating the counts of each query over thousands of servers. As the same query is often issued several times by multiple users across an entire month of search logs, we summarize each query with four ratios–search results clicks:overall clicks, ad clicks:overall clicks, spelling suggestion clicks:overall clicks, and query suggestion clicks:overall clicks.

A couple of steps were taken to ensure reliable ratios. We are classifying types, not tokens, and so limited ourselves to those queries with 100 or more clicks. This still leaves us with over half a million distinct queries for training and for testing, yet allows us to use click ratios as a substitute for annotating these huge data sets. If a query was only issued once and the user clicked on an ad, that may be more a reflection of the user, rather than reflecting that the query is 100% commercial. In addition, the ratios compare clicks of one type with clicks of another, rather than comparing clicks with impressions. There is less risk of a failure to find fallacy if we count events (clicks) instead of non-events (non-clicks). There are many reasons for non-clicks, only some of which tell us about the meaning of the query. There are bots that crawl pages and never click. Many links can't be seen (e.g., if they are below the fold).

Queries are labeled as positive examples of commercial intent if their ratio is in the top half of the training set, and negative otherwise. A similar procedure is used to label queries with the three other intents.

Our task is to predict future click patterns based on past click patterns. Note that a query may appear in both the test set and the training set, although not necessarily with the same label. In fact, because of the robustness requirement of 100+ clicks, many queries appear in both sets; 506,369 out of 591,122 of the test queries were also present in the training month. The overlap reflects natural processes on the web, with a long tail (of queries that will never be seen again) and a big fat head (of queries that come up again and again). Throwing away the overlap would both drastically reduce the size of the data and make the problem less realistic for a commercial application.

We therefore report results on various training set sizes so that we can show both: (a) the ability of the proposed method to generalize to unseen queries, and (b) the high performance of the baselines in a realistic setting. We vary the number of new queries by training the methods on subsets of 20%, 40%, 60%, 80%, and 100% of the positive examples (along with all the negative examples) in the training set. This led to the test set having 17%, 34%, 52%, 67%, and 86% actual overlap of these queries, respectively, with the training sets.

Click Type (Area on Results Page)	Query Type (Intent)	Example
Spelling Suggestion	Typo	www.lastmintue.com.au
Ad	Commercial Intent	ebay official
Query Suggestion	Suggestible	sears employees (where there are some popular query suggestions indicating how current employees can navigate to the benefits site, as well as how others can apply for employment)
Search Result	Standard Search	craigslist, denver, co

Table 1: Queries with a high percentage of clicks in each category

4 Three CQI Methods

4.1 Method 1: Look-up Baseline

The baseline method checks if a query was present in the training set, and if so, outputs the label from the training set. If the query was not present, it backs off to the appropriate default label: "non-commercial" for the commercial intent task (and "non-suggestible", "not a typo", etc. for the other CQI tasks). This very simple baseline method is effective because the ratios tend to be fairly stable from one month to the next. The query, *ebay official*, for example, has relatively high ad clicks in both the training month as well as the test month. The next section will propose an alternative method to address the main weakness of the baseline method, the inability to generalize beyond the queries in the training set.

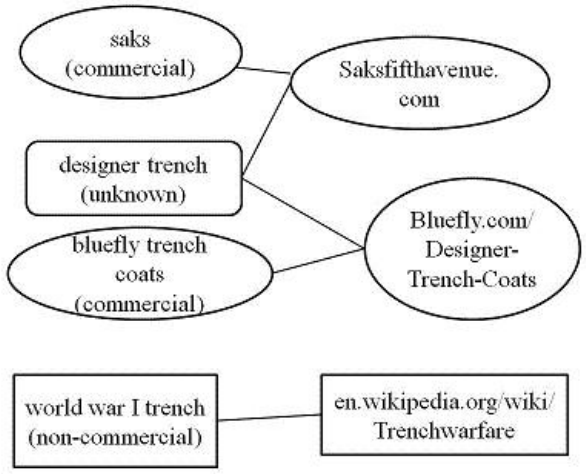


Figure 2: *saks* and *bluefly trench coats* are known to be commercial, while *world war I trench* is known to be non-commercial. What about *designer trench*? We can classify it as commercial because it shares URLs with the known commercial queries.

4.2 Method 2: Using Click Graph Context to Generalize Beyond the Queries in the Training Set

To address the generalization concern, we propose a method inspired by Yarowsky (1994). Word sense disambiguation is a classic problem in natural language processing. Some words have multiple senses; for instance, *bank* can either mean a riverbank or a financial institution, and for various tasks such as information retrieval, parsing, or information extraction, it is useful to be able to differentiate between the possible meanings.

When a word is being used in each sense, it tends to appear in a different context. For example, if the word *muddy* is nearby *bank*, the author is probably using the riverbank sense of the term, while if the word *deposit* is nearby, the word is probably being used with the financial sense.

Yarowksy (1995) thus creates a list of each possible context, sorted by how strong the evidence is for a particular sense. To classify a new example, Yarowsky (1994) finds the most informative collocation pattern that applies to the test example.

In this work, rather than using the surrounding words as context as in text classification, we consider the surrounding URLs in the click graph as context. A sample portion of the click graph is shown in figure 2. The figure shows queries on the left and URLs on the right. The click graph was computed on a very large sample of logs computed well before the training period. There is an edge from a query q to a URL u if at least 10 users issued q and then clicked on u.

For each URL, we look at its neighboring queries and calculate the log likelihood ratio of their labels in the training set. We classify a new query q according to URL^*, the neighboring URL with the strongest opinion (highest absolute value of the log likelihood ratio). That is, we compute URL^* with:

$$\text{argmax}_{U_i \in Nbr(q)} \left| \log \frac{\Pr(Intent|U_i)}{\Pr(\neg Intent|U_i)} \right|$$

If the neighboring opinion is positive (that is, $\Pr(Intent|URL^*) > \Pr(\neg Intent|URL^*)$), then the query q is assigned a positive label. Otherwise, q is assigned a negative label.

In Figure 2, we classify *designer trench* as a commercial query based on the neighbor with the strongest opinion. In this case, there was a tie between two neighbors with equally strong opinions: *www.saksfifthavenue.com* and *www.bluefly.com/Designer-Trench-Coats*. Both neighbors are associated with queries that were labeled commercial in the training set: *saks* and *bluefly trench coats*, respectively.

This method allows the labels of training set queries to propagate through the URLs to new test set queries.

4.3 Method 3: Hybrid ("Better Together")

We recommend a hybrid of the two methods:

- Method 1: the look-up baseline
- Method 2: use click graph context to generalize beyond the queries in the training set

Method 1 is designed for precision and method 2 is designed for recall. The hybrid uses method 1 when applicable, and otherwise, backs off to method 2.

5 Results

5.1 Commercial Intent

Table 2 and Figures 3(a) and 3(b) compare the performance on the proposed hybrid method with the baseline. When there is plenty of training material, both methods perform about equally well (the look-up baseline has an F-score of 84.1%, compared with the hybrid method's F-score of 85.3%), but generalization becomes important when training data is severely limited. Figure 3(a) shows that the proposed method does no harm and might even help a little when there is plenty of training data. The hybrid's main benefit is generalization to queries beyond the training set. If we severely limit the size of the training set to just 20% of the month, as in Figure 3(b), then the proposed hybrid method is substantially better than the baseline. In this case, the proposed hybrid method's F-score is 65.8%, compared with the look-up method's F-score of 28.4%.

5.2 Other types of clicks

Table 3 and Figures 4(a) and 4(b) show a similar pattern for the query suggestion task. In fact, the pattern is perhaps even stronger for the query suggestion task than commercial intent. When the full training set is used, the hybrid method achieves an F-score of 91.9% (precision = 91.5%, recall = 92.3%). When only 20% of the training data is used, the hybrid method has an F-score of 73.9%, compared with the baseline's F-score of 29.6%. A similar pattern was observed for clicks on search results.

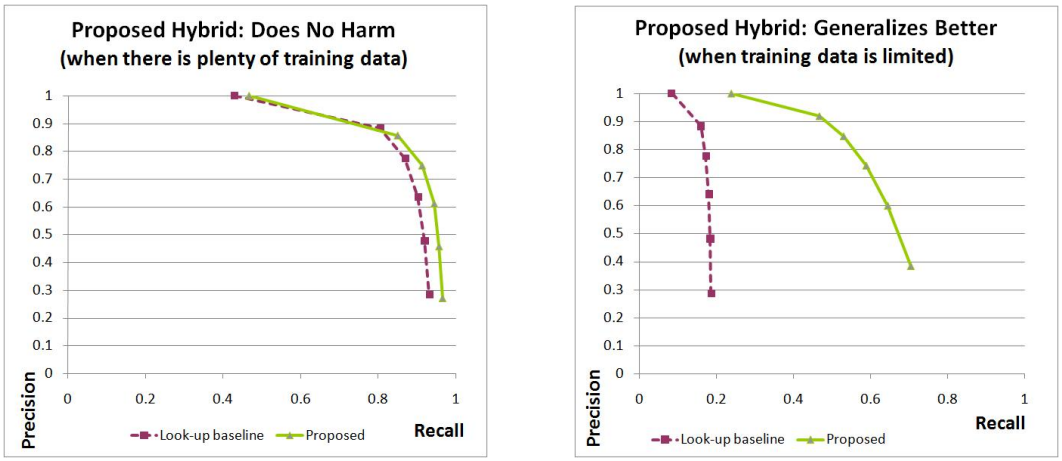
The one exception is the spelling suggestion task, where the context heuristic proved ineffective, for reasons that should not be surprising in retrospect. Click graph distance is an effective heuristic for many intents, but not for typos. Users who issue misspelled the query have the same goals as users who correctly spell the query, so we shouldn't expect URLs to be able to differentiate them. For misspelled queries, for example, *yuotube*, there are correctly spelled queries, like *youtube*, with the same intent that will tend to be associated with the same set of URLs (such as *www.youtube.com*).

6 Conclusion and Future Work

We would like to be able to distinguish web queries by intent. Unfortunately, we don't have annotated data for query intent, but we do have access to large quantities of click logs. The logs distinguish four types of clicks: (a) search results, (b) ads, (c) spelling suggestions and (d) query suggestions. Clicks on ads are evidence for commercial intent; other types of clicks are evidence for other intents. Click logs are huge sources of data, and while there are privacy concerns, anonymized logs are beginning to be released for research purposes (Craswell et al., 2009).

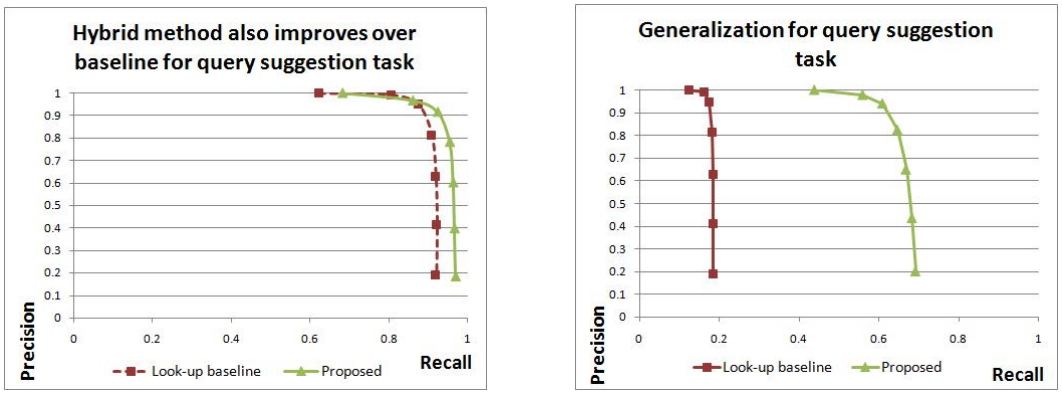
Besides commercial intent, queries can also be divided into two broader classes: queries in which the user is browsing and queries for which the user is navigating. Clicks on the ads and query suggestions indicate that users are browsing and willing to look at these alternative suggestions, while clicks on the search results indicate that the users were navigating to what they were searching for. Clicks on typos indicate neither, as presumably the users are not entering typos on purpose.

Just as dialogue management systems learn policies for when to allow *user* initiative (the user



(a) (b)

Figure 3: Better together: proposed hybrid is no worse than baseline (left) and generalizes better to unseen tail queries (right). The two panels are the same, except that the training set was reduced on the right to test generalization error.



(a) (b)

Figure 4: Similar to Figures 3(a) and 3(b), adding the decision list method generalizes over the look-up method for the "suggestible" task.

can respond in an open way) versus *system* initiative (the system asks the user questions with a restricted set of possible answers) (Relaño et al., 1999; Scheffler and Young, 2002; Singh et al., 2002), search engines may want to learn policies for when the user just wants the search results or when the user is open to suggestions. When users want help (they want the search engine to suggest results), more space on the page should be devoted to the ads and the query suggestions. When the users know what it is they want, more of the page should be given to the search results they asked for.

We started with a simple baseline for predicting click location that had great precision, but didn't generalize well beyond the queries in the training set. To improve recall, we proposed a context heuristic that backs off in the click graph. The backoff method is similar to Yarowsky's Word Sense Disambiguation method, except that context is defined in terms of URLs nearby in click graph distance, as opposed to words nearby in the text.

Our third method, a hybrid of the baseline method and the backoff method, is the strongest baseline we have come up with. The evaluation showed that the hybrid does no harm when there is plenty of training data, and generalizes better when there isn't.

A direction for further research would be to see if propagating query intent through URLs that are not direct neighbors but are further away, perhaps through random walk methods (Baluja et al., 2008;

Training Size	F-score			Precision / Recall		
	Baseline	Method 2	Hybrid	Baseline	Method 2	Hybrid
100%	84.1	75.6	**85.3**	88.2 / 80.4	76.6 / 74.6	85.7 / 85.0
80%	74.4	74.8	**83.5**	88.2 / 64.3	79.3 / 70.7	86.7 / 80.6
60%	62.4	72.9	**80.7**	88.3 / 48.2	82.5 / 65.3	87.9 / 74.6
40%	47.9	70.1	**76.0**	77.5 / 34.7	78.5 / 63.3	80.7 / 66.0
20%	28.4	62.5	**65.8**	77.6 / 17.4	75.9 / 53.1	74.3 / 59.1

Table 2: The baseline and hybrid methods have comparable F-scores when there is plenty of training data, but generalization becomes important when training data is severely limited. The proposed hybrid method generalizes better as indicated by the widening gap in F-scores with smaller and smaller training sets.

Training Size	F-score			Precision / Recall		
	Baseline	Method 2	Hybrid	Baseline	Method 2	Hybrid
100%	91.0	86.2	**91.9**	94.9 / 87.4	90.7 / 82.3	91.5 / 92.3
80%	80.5	85.2	**90.6**	94.9 / 69.9	91.6 / 79.7	91.9 / 89.4
60%	67.6	83.3	**88.6**	94.9 / 52.4	92.6 / 75.8	92.3 / 85.1
40%	51.0	79.5	**84.7**	94.9 / 34.9	87.6 / 72.7	93.0 / 77.8
20%	29.6	69.8	**73.9**	81.5 / 18.1	90.6 / 56.8	94.0 / 60.8

Table 3: F-scores on the query suggestion task. As in the commercial intent task, the proposed hybrid method does no harm when there is plenty of training data, but generalizes better when training data is severely limited.

Antonellis et al., 2008) improves classification.

Similar methods could be applied in future work to many other applications such labeling queries and URLs by: language, market, location, time, intended for a search vertical (such as medicine, recipes), intended for a type of answer (maps, pictures), as well as inappropriate intent (porn, spam).

In addition to click type, there are many other features in the logs that could prove useful for classifying queries by intent, e.g., who issued the query, when and where. Similar methods could also be used to personalize search (Teevan et al., 2008); for queries that mean different things to different people, the Yarowsky method could be applied to variables such as user, time and place, so the results reflect what a particular user intended in a particular context.

7 Acknowledgments

We thank Sue Dumais for her helpful comments on an early draft of this work. We would also like to thank the members of the Text Mining, Search, and Navigation (TMSN) group at Microsoft Research for useful discussions and the anonymous reviewers for their helpful comments.

References

I. Antonellis, H. Garcia-Molina, and C.C. Chang. 2008. Simrank++: query rewriting through link analysis of the clickgraph (poster). *WWW*.

S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. 2008. Video suggestion and discovery for youtube: taking random walks through the view graph. *WWW*.

D. Beeferman and A. Berger. 2000. Agglomerative clustering of a search engine query log. In *SIGKDD*, pages 407–416.

S.M. Beitzel, E.C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. 2004. Hourly analysis of a very large topically categorized web query log. *SIGIR*, pages 321–328.

S.M. Beitzel, E.C. Jensen, O. Frieder, D.D. Lewis, A. Chowdhury, and A. Kolcz. 2005. Improving automatic query classification via semi-supervised learning. *ICDM*, pages 42–49.

A.Z. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. 2007. Robust classification of rare queries using web knowledge. *SIGIR*, pages 231–238.

A. Broder. 2002. A taxonomy of web search. *SIGIR*, 36(2).

R. Chaiken, B. Jenkins, P.Å. Larson, B. Ramsey, D. Shakib, S. Weaver, and J. Zhou. 2008. SCOPE:

Easy and efficient parallel processing of massive data sets. *Proceedings of the VLDB Endowment archive*, 1(2):1265–1276.

D. Chakrabarti, D. Agarwal, and V. Josifovski. 2008. Contextual advertising by combining relevance with click feedback. *WWW*.

M. Ciaramita, V. Murdock, and V. Plachouras. 2008. Online learning from click data for sponsored search.

N. Craswell and M. Szummer. 2007. Random walks on the click graph. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 239–246.

N. Craswell, R. Jones, G. Dupret, and E. Viegas (Conference Chairs). 2009. Wscd '09: Proceedings of the 2009 workshop on web search click data.

H.K. Dai, L. Zhao, Z. Nie, J.R. Wen, L. Wang, and Y. Li. 2006. Detecting online commercial intention (OCI). *WWW*, pages 829–837.

J. Dean and S. Ghemawat. 2004. MapReduce: Simplified Data Processing on Large Clusters. *OSDI*, pages 137–149.

D. Downey, S. Dumais, D. Liebling, and E. Horvitz. 2008. Understanding the relationship between searchers' queries and information goals. In *CIKM*.

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

U. Lee, Z. Liu, and J. Cho. 2005. Automatic identification of user goals in Web search. In *WWW*, pages 391–400.

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

B. Piwowarski and H. Zaragoza. 2007. Predictive user click models based on click-through history. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 175–182.

J. Relaño, D. Tapias, M. Rodríguez, M. Charfuelán, and L. Hernández. 1999. Robust and flexible mixed-initiative dialogue for telephone services. In *Proceedings of EACL*.

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

K. Scheffler and S. Young. 2002. Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning. In *Proceedings of HLT*, pages 12–19.

S. Singh, D. Litman, M. Kearns, and M. Walker. 2002. Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. *Journal of Artificial Intelligence Research*, 16(1):105–133.

Bin Tan, Xuehua Shen, and ChengXiang Zhai. 2006. Mining long-term search history to improve search accuracy. pages 718–723. KDD.

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

Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan. 2004. Optimizing web search using web clickthrough data. In *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 118–126.

D. Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. *ACL*, pages 88–95.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *ACL*, pages 189–196.

D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Scholkopf. 2004. Learning with Local and Global Consistency. In *NIPS*.