

Exploiting User Search Sessions for the Semantic Categorization of Question-like Informational Search Queries

Alejandro Figueroa

Yahoo! Research Latin America
Av. Blanco Encalada 2120,
Santiago, Chile
afiguero@yahoo-inc.com

Günter Neumann

DFKI GmbH
Stuhlsatzenhausweg 3,
D-66123 Saarbrücken, Germany
neumann@dfki.de

Abstract

This work proposes to semantically classify question-like search queries (e.g., “*oil based heel creams*”) based on the context yielded by preceding search queries in the same user session. Our novel approach is promising as our initial results show that the classification accuracy improved in congruence with the number of previous queries used to model the question context.

1 Introduction

Open question answering (QA), i.e., fully automatic systems that find best answers to natural language questions of any type and domain, is still a challenging research problem. On the other hand side, search engines are getting smarter and smarter in order to fulfill users’ information requests. This motivates users to enter more sophisticated search queries (e.g., more complete questions) rather than few keywords, when they are looking for precise information needs (e.g., answers related to precise problems). This is also experienced by the fact that through search engines, it is likely to exploit answer databases of community based question answering (cQA) systems including Yahoo! Answers, if the search query is close to a QA-system like question. Then matching such a question with those in the cQA database is more likely to recognize plausible cQA phrases because of close textual relatedness. Furthermore, as the analysis of our data sources suggests, users often express semantically related series of questions in order to guide the search for better answers, and as such, are already performing interactions with search engines. In a general sense, searching is a sequence of queries in the same user session aimed at satisfying an underlying goal that the user is trying to achieve (Rose and Levinson, 2004).

Thus, we believe that it will be inevitable to further automatize a semantic analysis of search queries within user sessions, i.e., to analyze the semantic relatedness of a series of questions whether they constitute actually a session of semantically related questions entered by the same user.

Our contribution into these directions is the exploration of automatic methods to semantically classify question-like search queries, based on the context provided by preceding search queries in the same user session. An important aspect, tackled in this paper, is whether and how much contextual information extracted from user-specific search query sessions helps to effectively train and apply a model to predict the semantic category of a question-like informational search query (cf. (Broder, 2002; Rose and Levinson, 2004)).

Our method recognizes question-like queries by inspecting their associations with Yahoo! Answers pages via user clicks, providing the additional benefit of linking each query with an entry in the Yahoo! Answers category system. Thus our target semantic labeling set comprises 27 categories including business, environment, health, pets, sports and travel. As a consequence, we are able to completely automatize our approach without the need of manually annotated training material, and to automatically create a huge annotated corpus of semantically labeled question-like search queries. We then consider all search queries of a current session entered before the current labeled one as candidate sources for contextual information, and perform different experiments in order to explore the effect of different contextual window sizes. In a nutshell, our approach finished with 50.96% accuracy by exploiting nine previous search queries as window size.

2 Related Work

To the best of our knowledge, our work pioneers the idea of profiting from search sessions

for semantically categorizing question-like informational search queries. Broadly speaking, our study is related to community question answering (cQA) (Zhao et al., 2011), user session analysis (Cao et al., 2009), and closer to web query understanding (Reisinger and Pasca, 2011).

In a broad sense, (Rose and Levinson, 2004) proposed a framework for understanding the underlying goals of user searches. They outlined a taxonomy which its first level models three ends: informational (learn something by reading or viewing), navigational (going to a specific web-site) and resource (obtain videos, maps, etc.).

Later, in a more specific manner, the work of (Yin and Shah, 2010) seeks to understand search queries bearing a particular type of entity (e.g., musician) by classifying their generic user intents (e.g., songs, tickets, lyrics and mp3). They built a taxonomy of search intents by exploiting clustering algorithms, capturing words and phrases that frequently co-occur with entities in user queries, and by examining the click relationships between different intent phrases. Posteriorly, (Xue and Yin, 2011) extended this work by organizing query terms within named entity queries into topics, helping to better the understanding of major search intents about entities. The study of (Cheung and Li, 2012) presented an unsupervised approach to cluster queries with similar intents which, in their work, are patterns consisting of a sequence of semantic concepts or lexical items.

In effect, named entities cooperate on understanding user intents better, however detecting named entities in search queries is a difficult task, because named entities are not in standard form and search queries are typically very short (Guo et al., 2009). Thus, (Du et al., 2010) exploited query sequences in search sessions for dealing the lack of context in short queries, when distinguishing named entities on queries.

Our study focuses on the semantic categorization of question-like search queries, which cover a wide variety of informational queries that do not necessarily bear named entities. In particular, this paper studies the impact of preceding queries in user sessions for tackling the lack of context in this semantic categorization. Our approach is supervised trained with a large set of automatically tagged samples via inspecting click patterns between search queries and Yahoo! answers questions.

3 Our Approach

This section presents our automatic corpus acquisition and annotation technique, and later the features utilized by our supervised models.

3.1 Corpus Acquisition

Our corpus is distilled from a commercial search engine query log, more specifically, it considers queries in English submitted in the US from May 2011 to January 2013. We extracted ca. 65 millions full user sessions containing questions by keeping only those sessions connected to Yahoo! Answers via at least one user click. We assume that these clicks signal that, at some point during these sessions, users prompted questions and discovered pertinent information on the clicked Yahoo! Answer pages. Since sessions can cover a large period of time, and thus a wide variety of search needs, we split them into transactions by means of two criteria.

First, we benefited from the time difference that two consecutive queries were sent to the search engine. We used a gap of 300 seconds as session splitter, assuming that longer periods of time indicate that users are likely to have changed their search needs. This size for this temporal cut-off has been popularly used for segmenting query logs (Gayo-avello, 2009). Secondly, conventionally, navigational queries (e.g. , “twitter”) are prompted by users when they want to reach a particular web-site they bear in mind. As a rule of thumb, most frequent queries in search logs are navigational (Broder, 2002; Rose and Levinson, 2004). Thus we used all search queries having a frequency higher than 1,000 across our session corpus as additional transaction splitters.

Next, in order to study the impact of preceding queries in the session on the tagging of a new submitted question-like search query, we kept only transactions containing at least ten queries, where a user click links the tenth or a later query with Yahoo! Answers, and hence with one of its categories. In other words, we studied the impact of until nine historical queries. In total, this pre-processing gave us 1,098,778 transactions, where 15.87% and 3.41% of them are composed exactly of ten and 20 queries, respectively.

Table 1 shows a transaction consisting of 13 queries. Several ten-element transactions can be derived from one transaction. In this table, two query sequences: 1-10 and 3-12 are acquired,

Number	Search query	Clicked hosts
1	you tube how do i make a heel strap	Beauty & Style
2	cracked heel repair	
3	wraps for cracked heel repair	www.pantryspa.com
4	oil based moisturizer brands	
5	oil based moisturizer cream brands	ezinearticles.com
6	oil based moisturizer cream brands	www.alibaba.com
7	oil based moisturizer heel cream brands	www.amazon.com
8	oil based moisturizer heel cream brands	
9	oil based heel cream	
10	is vaseline considered a oil based moisturizer	Beauty & Style
11	vaseline uses	www.ehow.com
12	is vaseline an oil moisturizer	Beauty & Style
13	goodle	www.google.com

Table 1: A transaction (categories are shown for clicked Yahoo! Answers pages).

since queries ten and twelve are connected to Yahoo! Answers. Overall, we obtained 1,772,696 smaller transactions containing only ten elements, in which the 10th query is related to Yahoo! Answers by means of a user click.

3.2 Features

Basically, we took into account several features, which were a) derived from all search queries in the transaction; and b) targeted at inferring categories of preceding queries in the transaction, that is to say expect from the one being classified. In the first group, we have:

- **Bag-of-Words (BoW)** models a search query by their words and their respective frequencies.
- WordNet¹ semantic relations for extending search queries with a) words that include query terms in their the semantic range; and b) words that are included in the semantic range of any query term. The former (SR-A) comprises relations such as hypernyms (e.g., pressure → distress) and holonyms (e.g., professor → staff), while the latter (SR-B) relations like hyponyms (e.g., pressure → oil/gas pressure) and meronyms (e.g., service → supplication).

We only considered elements with an absolute frequency higher than three in the corpus. In the second group, that is attributes extracted exclusively from the window size of until nine search queries, we benefited from:

- **Clicked hosts (CH)** are pairs host/click count corresponding to previously clicked URLs

(see table 1). Note that a search query can be connected not necessarily with only one clicked host, but with many.

- **Category terms in URLs (CTU)** checks as to whether or not any of the terms in any previously clicked URL is a term in any of the categories in the Yahoo! Answers taxonomy. We use simple sign matchings to detect word boundaries within full URLs (e.g., slash, hyphen and underscore). We used lower-case for these matchings.
- **Yahoo! Answers Categories (YAC)** of previously clicked Yahoo! answers pages in the session. In our working example (see table 1), the category “*Beauty & Style*”.
- Similarly to YAC, we add words belonging to categories of previously **clicked Wikipedia pages (WC)**. We used words instead of full category names as many are not standardized.

4 Experiments and Results

In our empirical setting, we profited from SVM Multiclass as a multi-class classifier² (Crammer and Singer, 2001; Tsochantaridis et al., 2004). In all experiments, we use three-fold cross validation operating on our automatically annotated ten queries transaction corpus, since this collection is relatively large.

As for a **baseline**, we built a centroid vector (CV) for each class, and assign to each testing sample the label pertaining to the best scoring centroid vector afterwards. Here, we also conducted a three-fold cross-validation. Results achieved by this baseline and most SVM configurations indicate that the performance improves in tandem with

¹wordnet.princeton.edu

²svmlight.joachims.org/svm_multiclass.html

h	CV	SVM BoW	SVM BoW +						Combined
			CH	CTU	YAC	WC	SR-A	SR-B	
0	24.31	30.52	-	-	-	-	35.65	41.09	40.23
1	28.19	34.73	28.62	34.78	36.53	33.48	40.72	43.44	44.06
2	30.45	37.99	27.38	36.61	41.27	36.11	41.43	46.56	45.54
3	31.81	41.13	27.64	41.13	45.04	41.16	43.43	46.79	45.45
4	32.60	42.52	30.92	42.37	47.24	42.42	43.52	47.30	46.49
5	33.21	43.75	33.85	43.75	48.95	43.75	44.84	47.60	47.60
6	33.62	44.60	35.60	44.60	49.14	45.12	44.93	47.90	48.76
7	33.87	43.28	37.90	43.20	49.35	43.22	45.79	47.91	49.62
8	34.07	43.59	38.00	43.70	48.02	44.23	46.18	48.94	50.38
9	34.27	43.69	38.39	43.93	50.02	44.83	46.38	49.39	50.96

Table 2: Classification accuracy (%). h denotes the window (context) size.

the window size, that is the amount of session context. This comparison also shows that SVM exploits the context more efficiently: it requires a smaller number (6) of previous queries to accomplish a growth from 34.27% to 44.60% accuracy (see table 2). This is a key observation as it is also key to maximize the performance using as few as possible context, since this is not always available, especially when the user session is beginning.

Results reaped by models, that ignore context information ($h=0$ /"Combined" in table 2), show that features, attempting at discovering semantic hints about the new question-like search query, play a vital role. A combination of SR-A and SR-B improve the accuracy by about 10% (from 30.52% to 40.23%). This sheds light on the reason why the clicked host (CH) property was detrimental as several hosts (e.g., Wikipedia) are ambiguous, in other words, they aim at many potential categories. In fact, using this clicked host attribute the performance drops closer to the baseline.

Conversely, evidence from categories related to previously clicked Wikipedia (WC) links aids in enhancing the accuracy with respect to SVM+BoW (45.12% and $h=6$). This improvement is slight as the amount of clicked Wikipedia links is small with respect to the whole collection. On the other hand, categories of previously clicked Yahoo! Answers pages bettered the performance substantially (50.02% and $h=9$). A reason to this is the fact that we are dealing with question-oriented transactions, and hence clicks to Yahoo! Answers can be more frequent and relevant than clicks to Wikipedia. This finding indicates that specialized click patterns manifest across question-oriented search query transactions.

In light of our outcomes, we can conclude that semantic relations provided by WordNet at the word level are extremely useful. In particu-

lar, our figures show that adding SR-B type relations brought about an increase in accuracy from 30.52% to 41.09% and 49.39% without and with session context information, respectively.

Overall, our session context-aware approach combined (column "Combined" in table 2) with our features aimed at inferring semantic content (i.e., SR-A and SR-B) and query categories (i.e., CTU, WC and YAC) finished best (50.96%). This doubled the centroid vector baseline lacking of contextual information and it substantially improved a naive SVM built on BoW.

On a final note, inspecting the confusion matrix corresponding to the best configuration, we discovered that most recurrent misclassifications are due to categories "*Education & Reference*" and "*Health*", which were perceived as "*Science & Mathematics*". These error rates were (59.89%) and (36.25%), respectively.

5 Conclusions and Future Work

This study shows that the context provided by preceding queries in user search sessions improves the semantic labeling of QA-like informational search queries. Our results also point out to the positive contribution of semantically-based features.

As future work, we envision the use of linked data for drawing additional semantic inferences, thus assisting in improving the semantic tagging. Additionally, we envisage the use of sharper session segmentation techniques for identifying question-oriented transactions more accurately.

In principle, it would also be possible to build classifiers for checking as to whether or not a user input is a question-like search query, and for determining their semantic classes by some semantic database (e.g., an ontology). Actually, we also leave this open for future research.

References

- A. Broder. 2002. A Taxonomy of Web Search. In *SIGIR Forum 36:3-10*.
- Huanhuan Cao, Derek Hao Hu, Dou Shen, Daxin Jiang, Jian tao Sun, Enhong Chen, and Qiang Yang. 2009. Context-aware query classification. In *Research and Development in Information Retrieval*, pages 3–10.
- Jackie Chi Kit Cheung and Xiao Li. 2012. Sequence clustering and labeling for unsupervised query intent discovery. pages 383–392.
- Koby Crammer and Yoram Singer. 2001. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *Journal of Machine Learning Research*, 2:265–292.
- J. Du, Z. Zhang, J. Yan, Y. Cui, and Z. Chen. 2010. Using search session context for named entity recognition in query. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR*, pages 765–772.
- Daniel Gayo-avello. 2009. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179:1822–1843.
- Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Research and Development in Information Retrieval*, pages 267–274.
- Joseph Reisinger and Marius Pasca. 2011. Fine-grained class label markup of search queries. In *ACL 2011, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 1200–1209.
- D. E. Rose and D. Levinson. 2004. Understanding user goals in web search. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 13–19.
- Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *International Conference on Machine Learning*.
- Xiaobing Xue and Xiaoxin Yin. 2011. Topic modeling for named entity queries. pages 2009–2012.
- Xiaoxin Yin and Sarthak Shah. 2010. Building taxonomy of web search intents for name entity queries. In *World Wide Web Conference Series*, pages 1001–1010.
- Shiqi Zhao, Haifeng Wang, Chao Li, Ting Liu, and Yi Guan. 2011. Automatically generating questions from queries for community-based question answering. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 929–937.