2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)

A Clicked-URL Feature for Transactional Query Identification

Yingcheng Sun
Case Western Reserve University
Cleveland, OH, USA
yxs489@case.edu

Kenneth Loparo
Case Western Reserve University
Cleveland, OH, USA
kal4@case.edu

***Abstract*—Understanding query intents can help search engines to effectively improve their search quality. Click-through data has proven to be a valuable resource for query classification. In this paper, we propose a novel Clicked-URL (CURL) feature that uses semantic information extracted from user clicked URLs in the search results, and compare with the "key word list" to identify transnational query type. Experiments show that we can obtain relatively high accuracy in transnational query identification with CURL, and achieve an improved performance in query classification combing with other features.**

Keywords- search intent; transactional query identification; clicked URL feature

I. INTRODUCTION

Users' search intents can be basically classified into three categories according to the user goals: informational, navigational and transactional [1]. Understanding the underlying search intent associated with a query may effectively increase the level of search personalization [2, 3], and thus improve the search quality. For example, if a query is known to be transactional, we can improve search results by developing a special ranking function to meet users' transactional needs. Automatic query classification is usually performed by representing the queries with features extracted from the query itself and search engine logs. A bunch of features are proposed by pervious researchers, like linguistic-based properties [4], n Clicks Satisfied (nCS) Evidence, n Results Satisfied (nRS) Evidence [5], and etc.

Queries with different types of search intents exhibit different characteristics. In navigational queries, the user is interested in reaching a specific web site like "GitHub", "Harvard business school" and etc. In informational queries, the user does not have a particular page in mind and intends to learn more about a specific topic, such as ''amazon forest", "American civil war" and etc. In transactional queries, also known as resource queries, the user's objective is to obtain specific resource, not to learn some information but just to use the resource itself, and examples of such services are the "download of software"," buy flight ticket", and etc [7]. With analysis into search engine click-through data, we found that the URL addresses user clicked in the search results can be used as cue expressions for identifying transactional queries. We thus propose a Clicked-URL feature (CURL), that uses semantic information contained in hyperlinks to identify transnational search intents.

II. CLICKED-URL FEATURE DESCRIPTION

Query logs record the URL information every user clicked. For navigational queries, key words of queries often appear in the URL addresses because websites tend to use their names in the URL addresses, so Herrera et al. [6] proposed a method to identify navigational queries by calculating the percentage of retrieval items matched in URL addresses. However, the returned URL addresses for transactional queries may not contain the key words exactly extracted from queries but usually include semantic snippets reflecting users' transactional search needs. Table I lists five types of such snippets.

TABLE I. KEY WORD DESCRIPTION OF CLICKED URLS FOR TRANSACTIONAL QUERIES

Type	Type Description	Example
File	URL including words related with documents, zip files, games ,download tools, etc.	doc, pdf , zip, gz, book , soft, game, download, downloads, software, programs, cheat, cheats, exe, 7z, rar, tar.gz, cpp, dll, ttf, xml, xlsx, doc, docx
Video	URL including words related with video, movie, video playing websites or tools, etc.	video, playlist, show, tv, movie, film, mp4, wmv, m4v, mov, asf, avi, flv, wav
Music	URL including words related with music	sing, song, music, Music, mp3, wav, wma, midi
Picture	URL including words related with pictures	jpg, gif, bmp, image, images, photos, pics, picture
Travel	URL including words related with buying tickets online	online, train, flight, flights, buy, ticket, map, trip

Users often click different websites with different types of search intents. For transactional search intents, there are five types of clicked URL addresses: file, video, music, picture and travel, each of them represents a user's specific transactional need. For example, the "file" type includes hyperlinks pointing to file download websites, representing users' software or document acquisition needs. The "video" type includes URL addresses with video format, representing users' video acquisition needs. Key words of URL addresses for transactional queries are very different with the other two types of queries. For informational or navigational queries, the clicked URL addresses usually include the real names of

978-1-7281-2607-4/19/$31.00 ©2019 IEEE
DOI 10.1109/COMPSAC.2019.00156

the target website like "Amazon", "Facebook" and etc. We thus count the frequency of URL that includes the key words in Table I in search session *s*:

$$FURL(s) = \frac{\#\,(URL\ of\ session\ s\ that\ includes\ the\ listed\ key\ words)}{\#\,(URLs\ of\ session\ s)}$$

FURL calculates the ratio of the number of URL addresses that match the clicked URL key words of transactional queries in a search session. For each search query, users may submit multiple times making a large number of sessions, so we need to figure out how many sessions are involved in the transactional search intents:

$$CURL\ (q) = \frac{\#\,(Session\ of\ search\ query\ q\ that\ FURL(s) > \gamma)}{\#\,(Sessions\ of\ search\ query\ q)}$$

where γ is the threshold. Only sessions with the ratio FURL larger than γ will be counted as qualified for transactional search sessions, and CURL computes the ratio of "qualified" sessions to all sessions of a search query q, which can be used as a simple but discriminative feature for classifying user search intents.

III. EXPERIMENT

The data we used are the search engine logs collected from a commercial search engine *AOL* in the U.S. market during March and May in 2006, and a popular Chinese search engine *Sogou* in June in 2008. Figueroa [4] annotated 60,000 queries of *AOL* data set with search intents[1], 1000 of them are randomly selected for our experiment. We also annotated 1,000 queries from *Sogou* dataset[2] manually. We first explore the F1 score of query classification using *CURL* with the change of threshold γ. Figure 1 shows the result.

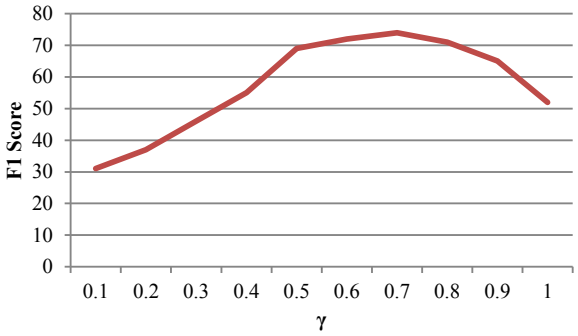

Figure 1. F1 score (%) of query classification with threshold γ.

In Figure 1, we can see that F1 score first increase with γ, and gets to the peak when γ is around 0.6 to 0.8, and then decreases. We set γ as 0.75 to make sure we can obtain the best classification performance. Next, we compare CURL with other kinds of features and combine them to prove the effectiveness. We use three other features: Bag of Words (Bow) as the linguistic feature that represents a web query as a term-frequency vector [4], n Clicks Satisfied (nCS) and n Results Satisfied (nRS) [5] as the click through features. Accuracy, recall and F1 score are calculated for different features and their combination. Table II shows the result.

[1] https://www.researchgate.net/publication/291184600_AnnotatedCorpus
[2] http://www.sogou.com/labs/resource/q.php

TABLE II. KEY WORD DESCRIPTION OF CLICKED URLS FOR TRANSACTIONAL QUERIES

Feature	Recall(%)	Accuracy(%)	F1(%)
Bow	65.9	68.4	67.1
nCS	64.5	67.1	65.8
nRS	67.2	70.3	68.7
CURL	**69.3**	**71.6**	**70.4**
Combination	72.1	73.9	73

From Table II, we can see that the query classification with CURL feature has higher accuracy and recall than other features, and obtains 70.4% of F1 score. It leads the best classification performance by combining all the four features that proves the effectiveness of CURL. The reasons that the intents of few queries are incorrectly classified are: 1) the intention type of some queries are ambiguous themselves [8], such as "New York City" representing informational needs for some people since they are interested in the introduction of this city but transactional needs for other people because they want to buy flight tickets to NYC. 2) For queries with complex semantic structures, the corresponded key words in clicked URL addresses are hard to find. It would be interesting to explore the methods to cope with the above issues to increase the accuracy in the future.

ACKNOWLEDGMENT

This work was supported by the Ohio Department of Higher Education, the Ohio Federal Research Network and the Wright State Applied Research Corporation under award WSARC-16-00530 (C4ISR: Human-Centered Big Data).

REFERENCES

[1] A. Broder, "A taxonomy of web search," in ACM Sigir forum, vol. 36, no. 2. ACM, 2002, pp. 3–10.

[2] Q. Li, Y. Zou, and Y. Sun, "User personalization mechanism in agentbased meta search engine," in Journal of Computational Information Systems, vol. 8. Springer, 2012, pp. 1–8.

[3] Q. Li, Y. Zou, and Y. Sun, "Ontology based user personalization mechanism in meta search engine," in 2012 2nd International Conference on Uncertainty Reasoning and Knowledge Engineering. IEEE, 2012, pp. 230–234.

[4] A. Figueroa, "Exploring effective features for recognizing the user intent behind web queries," Computers in Industry, vol. 68, pp. 162–169, 2015.

[5] Y. Liu, M. Zhang, L. Ru, and S. Ma, "Automatic query type identification based on click through information," in Asia Information Retrieval Symposium. Springer, 2006, pp. 593–600.

[6] Y. Lu, F. Peng, X. Li, and N. Ahmed, "Coupling feature selection and machine learning methods for navigational query identification," in Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006, pp. 682–689.

[7] Q. Li, Y. Sun, and B. Xue, "Complex query recognition based on dynamic learning mechanism," in Journal of Computational Information Systems, vol. 8. Springer, 2012, pp. 8333–8340.

[8] Q. Li and Y. Sun, "An agent based intelligent meta search engine," in International Conference on Web Information Systems and Mining. Springer, 2012, pp. 572–579.