ISSN 1674 – 3393

A Peer-reviewed International Scholarly Journal

中国文献情报(季刊)

# CHINESE JOURNAL OF LIBRARY AND INFORMATION SCIENCE

*(QUARTERLY)*

Volume 7 Number 4 2014
National Science Library, CAS

# Chinese Journal of Library and Information Science (CJLIS)

(Quarterly)

Sponsored by the Chinese Academy of Sciences

Volume 7 Number 4, December 25, 2014







**Published by:** National Science Library, Chinese Academy of Sciences
No. 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P.R. China
**Edited by:** Editorial Office of *Chinese Journal of Library and Information Science* (CJLIS)
No. 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P.R. China
**Printed by:** Beijing KEXIN Printing Co. Ltd., Beijing 102208, P.R. China. Tel: 86-10-62903036. Fax: 86-10-62805493
**Editor-in-Chief:** Prof. Xiaolin Zhang
**Typesetting:** Beijing Charlesworth Software Dev. Co. Ltd. (Beijing Modern Palace Building)
No. 20 Dongsanhuan, RD(South) , Chaoyang District, Beijing 100022, P.R. China. Tel: 86-10-67791601. Fax: 86-10-67799806.
**Distributed by:** Editorial Office of *Chinese Journal of Library and Information Science* (CJLIS)
No. 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P.R. China
**Subscription:** RMB ¥ 200/Issue, RMB ¥ 800/Volume domestically per year; US $ 199/Volume outside of China (including air shipping)
**Distributional Code** (邮发代号) 82-563
**Publishing Editor:** Lin PENG & Jing CAO

ISSN 1674 – 3393
CN 11-5670/G2

Research Paper

# Exploring features for automatic identification of news queries through query logs*

Xiaojuan ZHANG & Jian LI†

School of Computer and Information Science, Southwest University, Chongqing 400715, China

Received: Aug. 22, 2014
Revised: Feb. 6, 2015
Accepted: Feb. 14, 2015

## Abstract

**Purpose:** Existing researches of predicting queries with news intents have tried to extract the classification features from external knowledge bases, this paper tries to present how to apply features extracted from query logs for automatic identification of news queries without using any external resources.

**Design/methodology/approach:** First, we manually labeled 1,220 news queries from Sogou.com. Based on the analysis of these queries, we then identified three features of news queries in terms of query content, time of query occurrence and user click behavior. Afterwards, we used 12 effective features proposed in literature as baseline and conducted experiments based on the support vector machine (SVM) classifier. Finally, we compared the impacts of the features used in this paper on the identification of news queries.

**Findings:** Compared with baseline features, the *F*-score has been improved from 0.6414 to 0.8368 after the use of three newly-identified features, among which the burst point (*bst*) was the most effective while predicting news queries. In addition, query expression (*qes*) was more useful than query terms, and among the click behavior-based features, news URL was the most effective one.

**Research limitations:** Analyses based on features extracted from query logs might lead to produce limited results. Instead of short queries, the segmentation tool used in this study has been more widely applied for long texts.

**Practical implications:** The research will be helpful for general-purpose search engines to address search intents for news events.

**Originality/value:** Our approach provides a new and different perspective in recognizing queries with news intent without such large news corpora as blogs or Twitter.

**Keywords** Query intent; News query; News intent; Query classification; Automatic identification

* This work is supported by the Social Science Planning Foundation of Chongqing (Grant No.: 2011QNCB28).
† Corresponding author: Jian Li (E-mail: lijian@swu.edu.cn).

CJLIS
Vol. 7 No. 4, 2014
pp 31–45
National Science Library, Chinese Academy of Sciences
*http://www.chinalibraries.net*

Research Paper

## 1 Introduction

To better serve users who prefer to search news through general-purpose search engines, Web search engine developers have tried to aggregate specialized news corpora into Web search results. However, due to limited space on each search page, Koenig et al.[1] suggested to rank news pages for queries only with news intents.

Many researchers analyze blogs and Twitter to identify news queries, as keywords associated with news events are frequently found in current corpora of news and blog articles. For example, Koenig et al.[1] used query distribution in blogs, newswire and Wikipedia and estimated the click-through rate for news queries. Louis et al.[2] calculated the similarity between queries based on URLs, titles and abstracts of clicked news or blog pages and improved the quality of news query prediction. In combination with users' feedback, Diaz[3] applied click-through data of search engines for training a news query classifier. However, owing to the difficulties in obtaining large corpora of news, approaches with large data cannot be widely applied.

Since query logs contain vast amount of information about users' information needs and their search interests, researches of analyzing query logs have emerged and rapidly increased. According to Broder's[4] definition of query classification (informational, transactional and navigational), some researchers have proposed effective features[5,6] to automatically classify users' intent based on query logs, such as query length[7], query terms[8], *n* clicks satisfied (nCS)[9], top *n* results satisfied (nRS)[9], click entropy[10,11], click distribution[11], cPopular[12] and cDistinct[12]. Wu et al.[13] identified user intents based on features extracted from various sources including query logs and Web results retrieved, and they reported a best classification accuracy rate up to 88.5%.

Moreover, several studies tried to identify news queries via query logs. For instance, Maslov et al.[14] extracted queries related to real-life events from a general-purpose Web search engine, by using relative query log frequencies. They found that news-related queries were short in length but revealed such important aspects of events as locations, dates, actors or event types. McCreadie et al.[15] used Amazon's Truck Platform to label news queries; Hassan et al.[16] predicted queries with news intents by analyzing the relationship among place names contained in queries, the geographical locations of information users exposed from their IP addresses and their preference of clicking news Web pages. However, it is difficult to obtain users' IP addresses due to the concern of protection of users' privacy, and it is still unclear how to automatically identify news queries based on the annotation[14].

Compared with external knowledge resources, query logs are a very useful resource for studying users' information needs, because they can directly reflect people's search interests[17]. However, few studies have focused on the use of

National Science Library, Chinese Academy of Sciences

classification features[4] contained in the news query, and their impact on the quality of news query identification remains unknown. In this study, we intend to explore features to identify news queries through query logs, with an aim to provide a new insight into the way to address search intents of news events.

## 2 Methodology

### 2.1 Data source

All data were collected from the Sogou query logs of Sogou.com[①]. The query log used in this study covered a one-month period (from June 1, 2008 to June 30, 2008) and consisted of 20 million records issued by more than 657,000 users. As shown in Table 1, each record contains the following six items: 1) The query time (the time the query was issued), 2) a user ID, 3) a query, 4) the item rank (the rank of clicked result), 5) click sequences (the number in the click sequence in a search session), and 6) the clicked URL.

**Table 1 Dataset examples of the Sogou query log**

| Query time | User ID | Query | Item rank | Click sequence | Clicked URL |
|---|---|---|---|---|---|
| 00:00:03 | 8234353 | TV program download | 9 | 6 | www.tvbsale.com/ |
| 00:00:04 | 720986435 | You HaNa | 2 | 3 | club.koook.com/9226 |

After the removal of the queries of Web URLs, adult contents and single stop words, the time-based method was used to segment sessions, and time-out threshold was set to be 15 minutes[18]. In order to make our sample more representative and unbiased, we extracted queries from June 11 to June 20, 2008, i.e., the intermediate time period of June 2008 in this experiment, then used the Poisson sampling strategy proposed by McCreadie et al.[15], finally, we got a total of 11,068 sample queries.

### 2.2 Manual labeling of news queries

News queries used in this study were defined according to Louis et al.[2]. However, it is of great challenge to decide whether a query is news-related. For example, the query "Wenchuan" is more likely to be news-related after the outbreak of the Wenchuan Earthquake on May 12, 2008, before that, it is more associated with such non-news type of information as geography and travel, etc.

All the queries were marked by 9 labelers. If a word appeared in news titles, abstracts or body fields and a page that contained the word was created on the day of the query term, it was marked as a query with a news intent.

① Sogou is the first general-purpose Chinese-language search engine launched by Sohu Co. in 2007, and it has a collection of Web pages up to 10 billion. More details are available at http://www.sogou.com/labs/dl/q.html.

Considering that the frequently clicked URLs can express users' search purpose, we integrated the top clicked query news pages on the query issued day into the labeling interface, with an intention that the labelers can get more information about the query-related news event with Sogou search engine at that time. All annotated information was stored in MySQL② database.

Finally, we got a total of 1,220 news-related queries, which accounted for ca 11% of the total 11,068 sampled queries. This percentage was roughly in line with the finding reported in Bar-Ilan et al.'s study[19]. The annotator agreement was measured by Cohen's Kappa[20] and was calculated at 0.78, which means it has met the requirements of the substantial levels of agreement[21] and the labeling work is very effective.

## 3 Features for news queries

### 3.1 Features identified from sample data

#### 3.1.1 Content-based query feature

As the size and content of a query are good indicators for query analyses[22], news query features were investigated through the analysis of the query content. In our data sample, we found that the average length of both news and non-news queries was about 5 characters (Table 2) and the average number of terms in a query was about 2 words (Table 3).

**Table 2 Length of news and non-news queries in our datasets**

| Query length | Proportion of news query (%) | Proportion of non-news query (%) |
|---|---|---|
| With 1 character | 3 | 5 |
| With 2 characters | 4 | 5 |
| With 3 characters | 8 | 4 |
| With 4 characters | 23 | 19 |
| With 5 characters | 34 | 38 |
| With 6 characters | 20 | 23 |
| With more than 6 characters | 8 | 6 |

According to Maslov et al.[14], news-related queries were short in length, but could illustrate such important aspects of an event as locations and dates. In addition, we considered other two kinds of person named entities, i.e., person and organization, in our analysis and used ICTCLAS③ as our tool to segment query words, then we

② MySQL is a widely used open-source relational database management system (http://en.wikipedia.org/wiki/MySQL).

③ ICTCLAS, a Chinese Lexical Analysis System which has been developed by the Institute of Computing Technology of the Chinese Academy of Sciences. Available at http://www.ictclas.org/.

**Table 3 Query terms contained in queries**

| Query | Proportion of news query (%) | Proportion of non-news query (%) |
|---|---|---|
| With 1 query term | 18 | 23 |
| With 2 query terms | 42 | 39 |
| With 3 query terms | 31 | 26 |
| With more than 3 query terms | 9 | 12 |

recognized person, organization and place entities. The high accuracy was reported as 98.45%[23], and it was fast and easy to operate ICTCLAS.

In the whole labeled sets, there were only 8 queries containing time entities (e.g., "Olympics in 2008", "College Entrance Examination in 2008") and 10 queries containing organization entities (e.g., "Harbin Finance University", "Admission in Peking University"), which accounts for 0.15% and 0.2% of all queries, respectively. As displayed in Table 4, person entities and place entities were important features of news queries.

**Table 4 News and non-news queries with place or person named entities**

| Query | Frequency | Percentage (%) |
|---|---|---|
| News queries WITH a place or person name | 964 | 79 |
| News queries WITHOUT a place or person name | 256 | 21 |
| Non-news queries WITH a place or person name | 2,856 | 29 |
| Non-news queries WITHOUT a place or person name | 6,992 | 71 |

Since users often raise multiple queries and conduct multiple rounds of interaction with a search engine for specific information, and the context of their previous queries in the same session might help us understand the information behavior of users[17], thus, we regard the contextual information as another important clue of news queries. In our study, the contextual information refers to the remaining words in a query except name entities (place, person and organization). Table 5 outlines 5 most frequently used contextual words for news and non-news queries when considering entity type information in queries.

**Table 5 Occurrence of 5 top-ranked contextual words (events) in news and non-news queries**

| Query type | Top-ranked 5 contextual words | Percentage of the total query (%) |
|---|---|---|
| News queries WITH person entity | Scandal, marriage, events, death, divorcement | 81 |
| News queries WITH place entity | National entrance examination, earthquake, torch, incidents, scoring test papers | 72 |
| Non-news queries WITH person entity | Download, pictures, movie, resume, blog | 83 |
| Non-news queries WITH place entity | Education, food, bus station, entertainment, house prices | 86 |

Table 5 implies that different query categories contained different contextual words. In view of this, we calculated the capability of contextual words while representing news queries. As shown in Eq. (1), $P(w)$ represents the value of the contextual word $w$, $N_{\text{news}}(et, w)$ denotes the number of news queries which contain entity category $et$ as well as the query term $w$ in the news query training set, $N_{\text{news}}(w)$ is the total news query with the query term $w$, and $et$ the entity category. Analogously, $N_{\text{nonnews}}(et, w)$ means the total number of queries containing the query term $w$ and entity category $et$ in the non-news training set.

$$P(w) = \lambda \times \frac{N_{\text{news}}(et, w)}{N_{\text{news}}(et, w) + N_{\text{nonnews}}(et, w)} + (1 - \lambda) \times \frac{N(et, w)}{N_{\text{news}}(w) + N_{\text{nonnews}}(w) + 1} \quad (1)$$

The value of entity $et$ may be person or place named entity. If a query does not contain any entity, the value of $N_{\text{news}}(et, w)$ or that of $N_{\text{nonnews}}(et, w)$ is set to be 0. In our training sets, $\lambda$ is a weight set as 0.5 according to Hassan et al.'s study[16]. If a query contains words representing named entities, we regard these words as contextual words and calculate their values with Eq. (1). However, if a query contains more than one contextual word, only the word with the maximum $P(w)$ value will be considered.

### 3.1.2 Time-based query feature

According to Zhao et al.[24], if an event of something unique happens at a certain point of time, this particular interval can be regarded as the duration of a news event. Based on this definition, the duration will be treated as news query time if the news is issued at the time point when a query-related event occurs.

As reported by Sun & Hu[25], many users search for related information after a news event happens, which makes the news-related queries more popular, and the popularity can be reflected in the query logs[26]. As shown in Fig. 1, the horizontal axis represents the specific date in June, 2008, and the vertical is the ratio of the query frequency. The curve of news queries fluctuates a lot, suggesting a huge difference existing between the maximum and the minimum probability.

In Fig. 1, there is a smooth curve before the occurrence of the curve peak for news queries within a month period, which is followed by another smooth curve. In this paper, the duration of a query peak was regarded as that of the news event, so the peak apex was called a "query burst point". As indicated in Fig. 1(a) and Fig. 1 (b), the burst point of the news queries is on June 11 and 12, 2008, respectively.

In order to calculate the duration interval of a news event, for a specific query, we first calculated the maximum probability of its occurrence in the month, then

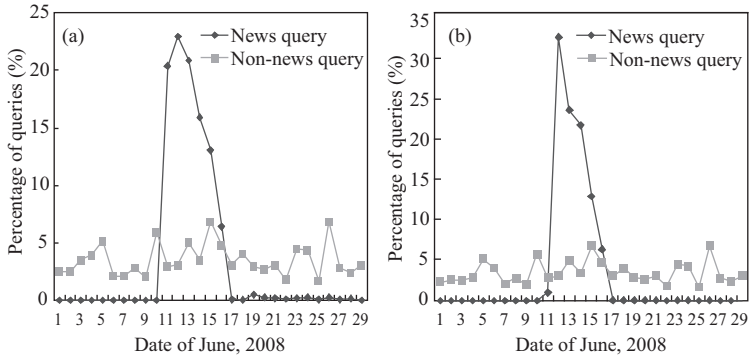


Fig. 1 Temporal distribution of news query examples and non-news query examples. (a) An actor named Lu You (news query example) and actors or actress change of the TV Series *New Dream of the Red Mansion* (non-news query example); (b) Ping-Pong videos (news query example) and human resource agencies in Shenzhen (non-news query example).

the difference between each date occurrence and the maximum probability. At last, we used the average difference to measure the curve volatility. To make a further analysis of news and non-news queries in the dataset, we divided the average deviation into 5 numerical intervals (Fig. 2). The ratio of the interval "<10%" for non-news query is 80%, indicating that the average deviation of approximately 80% of the non-news queries fell within this interval.

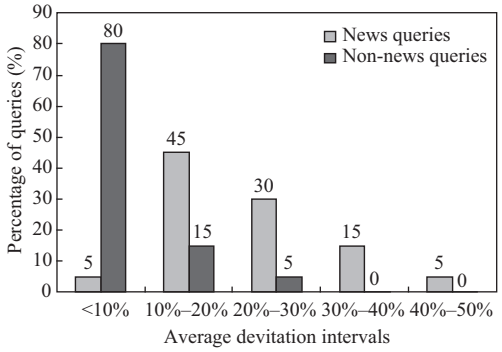


Fig. 2 Ratios of news or non-news queries in the average deviation intervals.

Figure 2 also shows that the average deviation of most non-news queries is less than 10% whereas that of most news queries is greater than 10%. This is also the reason why the query distribution in a month was used to predict the duration time of news events related to the query.

Suppose that the difference between the occurrence probability on the date $d_i$ and that of the date $d_{i-1}$ is greater than 10% in a month, the date $d_i$ is then defined as the query burst point. As seen in Fig. 1, news popularity decreases gradually, this means the volatility is lower before the final day of the news event. Based on this observation, we assumed that if the difference between the maximum probability and the probability on the day $d_{i+k}$ is greater than 10%, $d_{i+k}$ will be the news duration end point. The period of $k$ days between the news burst point and the end point is the duration time of the news event.

For the 1,220 news queries investigated in our study, we found the average duration time of news event was 5 days. This means, if a burst point exists during the first 4 days before $d_i$ for a specific given query, the query is more likely to have a news intent. In order to find whether a query had a burst point, we used Boolean value[④] to represent this feature. If a query has a burst point, its feature value is set to be 1; otherwise, it is 0.

### 3.1.3 Click behavior-based query feature

According to Gaugaz1 et al.[27], if a user's query contains a news intent, there must be various descriptions of the same topic on the clicked Web pages. In other words, topical similarities may exist among all search results. We thus selected the Web pages clicked for a given query with top-ranked 20 click frequencies in the first 4 days to analyze the topical similarities.

The data processing, such as Web page cleanups, text extraction and stop word removal were accomplished by our programming. Furthermore, Chinese word segmentation was performed by the Chinese word segmentation tool, ICTCLAS. Meanwhile, we applied this tool to recognize person and place entities in pages, and employed regular expression to identify time entities in queries.

In this paper, the topical similarity between two Web pages was calculated with Eq. (2), where $sim(a_1, a_2)$ denotes the content similarity between two Web pages. The higher the similarity, the more related their contents. The boost constant $boost(t)$ equals to 3 if the term is an entity (i.e., person name, place name or time); otherwise, it is 1.

$$sim(a_1, a_2) = \sum_{t \in a_1} [tf(t, a_2) \times idf(t) \times boost(t)] \quad (2)$$

In addition, the feature value $tf(t, a_2)$ denotes the ratio of a word $t$ to the total number of words contained in Web page $a_2$ and is calculated as Eq. (3), where $count(t \text{ in } a_2)$ refers to the frequency of the word $t$ and $size(a_2)$ the total terms in Web page $a_2$.



④ See http://en.wikipedia.org/wiki/Bóolean_data_type.

$$tf(t,a_2)=\frac{count(t \text{ in } a_2)}{size(a_2)} \quad (3)$$

Intuitively, a news event is often reported by different websites or Web pages, but its core content will remain unchanged. So words representing entities are always more important when calculating the similarity within two pages. In light of this, the inverse document frequency $idf(t)$ is calculated with Eq. (4), where $count(a_2 \text{ has } t)$ refers to the number of documents in Web page $a_2$ containing the word $t$.

$$idf(t)=1+\log\left[\frac{N}{count(a_2 \text{ has } t)}\right] \quad (4)$$

## 3.2 Features proposed in literature

### 3.2.1 Content-based query features

- Terms. As Ruocco & Ramampiaro[28] reported, heterogeneous tags play an important role in ranking and extracting hot-spot related tags. Such heterogeneous words as "earthquake" and "accident" are more likely to appear in news queries.
- Number of terms[28]. This feature is based on the idea that if a user wants to find some news related to a specific topic, he or she will construct a more specific query to express his or her information need. Therefore, news queries are assumed to be longer on average than non-news queries.

### 3.2.2 Time-based query features

- Average session length (*sl*)[29]. It refers to the average length of the sessions in which a query has been formulated. Our intuition is that news searchers use less time than non-news searchers.
- Average session time (*ast*)[25]. Claypool et al.[26] found that the time spent on a page has a strong correlation with users' explicit interest. On average, we assume that news searchers spend less time than non-news searchers in a query session.
- Query popularity (*qpop*). Namely, the frequency of query occurrence in a query log[27]. We assume that news queries are more popular than non-news queries in the interval of the outbreak of a news event.

### 3.2.3 Click behavior-based query features

- News URL(*nu*). The ratio of news URL clicks against the total clicks number of a query[1]. The intuition behind is that news queries are more likely to appear in news URL addresses.

- MedianClick(*mc*). Median of click distribution[9]. For a news query, the median of click distribution could highly skew to a particular document; while for a non-news query, the median of click distribution must be much larger since users click on multiple documents.
- cPopular(*cp*). Similar to the percentage of the most clicks for the query[12], queries associated with clicks of unique topics contain news intents.
- Click entropy(*ce*)[10]. Smaller click entropy means that the majority users click a limited number of Web pages, thus, when a user issues a news query, he/she is more likely to click on more often-visited Web pages.
- Domain ClickEntropy(*de*). Namely, click entropy based on domain clicks. According to Yuan et al.[10], users tend to click on a short domain name to retrieve Web pages, which would increase the click entropy and improve the identification accuracy. Therefore, we suppose users are more prone to click on more different domains.
- nCS[9]. That is, the percentage of query sessions less than *n* clicks. According to Jansen et al.[7], the default value *n*=2 was adopted in our study. We assume that news searchers are likely to click on more Web pages in each search session.
- nRS. Percentage of query sessions that only involve top *n* results[9]. According to Jansen et al.[7], the default value *n*=5 was set for this feature. We assume news query searchers are more likely to click on the top-ranked Web pages.

## 4 Experiments

### 4.1 Datasets

A total of 11,068 queries from Sogou query log were used to evaluate the effectiveness of all features in predicting news queries. We used the SVM$^{light}$ implementation software package[5] to classify queries. The 10-fold cross-validation method[6] was adopted in the experiments.

### 4.2 Evaluation indicators

We used precision, recall and *F*-score (*F*1) measures to determine the effectiveness of all features in news query identification. Precision *p* is the proportion of all correctly classified examples in the set assigned to the target class, and recall *r* the proportion of all correctly classified examples with the target class. *F*1 is a combination of precision and recall defined as $\frac{2 \times p \times r}{p + r}$ . These metrics were first separately calculated for news and non-new queries, then were averaged, i.e.,

[5] Available at http://www.svmlight.joachims.org/.
[6] See http://en.wikipedia.org/wiki/Cross-Validation_statistics.

macro-average. For all datasets, we used $C_{news}$, $C_{nonnews}$ and $C_{all}$ respectively to present the accuracy classification obtained for news, non-news and all query taxonomies, which implied two different tasks: Distinguishing news from non-news queries and calculating the accuracy of identified news and non-news queries.

### 4.3 Results

For support vector machine (SVM) classifier, three function kernels (linear, polynomial and radial basis function) were used to test their classification accuracy, respectively. Table 6 shows that the radial basis function (RBF) has a better accuracy over other kernels. Therefore, we used the SVM and RBF to accomplish the following experiments, and default settings were adopted for other parameters.

**Table 6 *F*1 values using different functions kernel of SVM**

| Function kernels | $C_{news}$ | $C_{nonnews}$ | $C_{all}$ |
|---|---|---|---|
| Linear | 0.7354 | 0.7753 | 0.7248 |
| Polynomial | 0.7612 | 0.7314 | 0.7400 |
| Radial basis function (RBF) | 0.8530 | 0.8012 | 0.8368 |

Moreover, the previously proposed features in Section 3.2.3 were used as baseline features. We employed the same classifier when the baseline features were used to determine news queries in the data sample. As shown in Table 7, the macro-averaged *F*-score of the classification result was 0.8368 when all features were used, i.e., 8 out of 10 news queries can be generally classified correctly. But the impacts of query expression (*qes*), query burst point (*bst*) and click results similarity (*crs*) on classification accuracy were evident since the macro-averaged *F*-score was improved from 0.6414 to 0.8368.

**Table 7 Comparison of classification results after adding the features of *qes*, *bst* and *crs***

| Methods | Macro-averaged precision | Macro-averaged recall | Macro-averaged *F*-score |
|---|---|---|---|
| Baseline features | 0.6453 | 0.6375 | 0.6414 |
| Adding the features *qes*, *bst* and *crs* | 0.8404 | 0.8332 | 0.8368 |

## 5 Discussions and conclusions

To study the impact of *qes*, *bst* and *crs* on classification accuracy of predicting news intent, we calculated their macro-averaged *F*-score respectively. Table 8 lists the results when each of 12 features in Section 3.3.2 and our newly identified features *qes*, *bst* and *crs* was removed.

As indicated in Table 8, all features except *mc* had a positive impact on improving the accuracy of classification. Although query-based features (such as *terms* and

**Table 8 Classification accuracy with one feature removed**

| Feature removed | Macro-averaged *F*-score | Feature removed | Macro-averaged *F*-score |
|---|---|---|---|
| *none** | 0.8368 | *cp* | 0.8321 |
| *terms* | 0.8298 | *ce* | 0.8326 |
| *#terms* | 0.8331 | *de* | 0.8321 |
| *qpop* | 0.8321 | *sl* | 0.8319 |
| *nu* | 0.8210 | *ast* | 0.8318 |
| *nCS* | 0.8359 | *qes* | 0.8221 |
| *nRS* | 0.8343 | *bst* | 0.8120 |
| *mc* | **0.8369** | *crs* | 0.8300 |

*Note: "none" means all features are used. Results higher than 0.8368 are shown in bold.

*qes*), click behavior-based features (such as *nCS*, *nRS*, *nu* and *cp*), and time-based features (such as *sl* and *qpop*) were all effective in identifying news queries, query burst point (*bst*) was found to be the most effective one. When *bst* was removed, the classification accuracy value was 0.8120, the smallest among all features. This indicates the usefulness of query burst point in predicting news queries.

In terms of query-based feature, query expression (*qes*) was more effective than query terms (*terms*), this is because *qes* can provide clues of news/non-news queries with both entity information and contextual words.

In addition, *qes* was more effective than clicked result similarity (*crs*), and this may be due to the reason that *qes* can describe users' search intents in a direct way, and the clicked results can contain noisy information which also affects the final results when calculating with *crs*.

Moreover, news URL (*nu*) was found to be the most useful feature among the click behavior-based features, which implies that the click information on news pages was more important than that on common documents. Furthermore, the removal of median of click distribution (*mc*) also led to a better accuracy. This is probably because the search goals of news and non-news queries were both likely to be informational, and numerous result pages were thus likely to be clicked.

However, we only used the features extracted from query logs in this study of news query identification. Log data alone present a partial view of user behavior and their interactions with the information retrieval systems only[17]. To better understand user's search intents, approaches such as one-by-one studies or questionnaire surveys are required urgently. Wu et al.[13] also indicated that the usage of different features extracted from diversified sources may help to improve the efficiency of identification of query intents, nevertheless, we need to prove if the classification accuracy can be improved when other source features except for query logs are used in the future study.

Furthermore, the annotation accuracy of recognizing named entities in queries still need be improved. Instead of short queries, ICTCLAS is effective for long texts

such as documents segmentation, and the impact of organization entities was not been considered on the identification of news queries in this study. In addition, the cost could be high if such features as query burst point be employed in the classification of news queries. For instance, some name or place entities could provide important clues for identifying a news query during a specific time period. Consequently, in order to apply this feature, editors of general-purposed search engines need have more training on how to label queries from time to time.

## Author contributions

X.J. Zhang (zhangxiaojuan624@gmail.com) designed the research and wrote the draft of the manuscript. J. Li (lijian@swu.edu.cn, corresponding author) conducted the experiments. Both authors have actively participated in the revision of the paper.

## References

1 Koenig, A.F., Gamon, M., & Wu, Q. Click-through prediction for news queries. Proceedings of SIGIR 2009. New York: ACM Press, 2009: 347– 354. Retrieved on August 29, 2014, from http://research.microsoft.com/apps/pubs/?id=80233.

2 Louis, A., Crestan, E., & Billawala, Y., et al. Use of query similarity for improving presentation of news verticals. In Brambilla, M., Casati, F., & Ceri, S. (Eds.) Proceedings of the 1st International Workshop on Searching and Integrating New Web Data Sources - Very Large Data Search, 2011: 1–6. Seattle. Retrieved on August 29, 2014, from http://ceur-ws.org/Vol-880/.

3 Diaz, F. Integration of news content into Web results. Proceedings of WSDM 2009. New York: ACM Press, 2009: 182–191. Retrieved on August 29, 2014, from http://dl.acm.org/citation.cfm?id=1498825&dl=ACM&coll=DL&CFID=547401920&CFTOKEN=13528633. DOI: 10.1145/1498759.14988254.

4 Broder, A. A taxonomy of Web search. SIGIR Forum, 2002, 36(2): 3–10. Retrieved on August 29, 2014, from www.sigir.org/forum/F2002/broder.pdf.

5 Brenes, D.J., Gayo-Avello, D., & Perez-Gonzalez, K. Survey and evaluation of query intent detecting methods. Proceedings of the 2009 Workshop on Web Search Click Data. New York: ACM Press, 2009: 1–7. Retrieved on August 29, 2 014, from http://65.54.113.26/Publication/6263853. DOI: 10.1145/1507509.1507510.

6 Lu, W., Zhou, H.X., & Zhang, X.J. The review of query intent. Journal of Library Science in China (in Chinese), 2013, 1: 103–114. Retrieved on February 1, 2015, from http://cn.oversea.cnki.net/kcms/detail/11.2746.G2.20121126.2007.001.html.

7 Jansen, B.J., Booth, D.L., & Spink, A. Determining the informational, navigational, and transactional intent of Web queries. Journal of Information Processing and Management, 2008, 44(3): 1251–1266. Retrieved on August 29, 2014, from http://dl.acm.org/citation.cfm?id=1351372. DOI: 10.1016/j.ipm.2007.07.015.

8 Kato, P.M, Yamamoto, T., & Ohshima, H., et al. Cognitive search intents hidden behind queries: A user study on query formulations. The 23rd ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2014: 313–314. Retrieved

on February 1, 2015, from http://dl.acm.org/citation.cfm?id=2577279&dl=ACM&coll=DL&CFID=623465122& CFTOKEN= 76345971. DOI: 10.1145/2567948.2577279.

9 Liu, Y.Q., Zhang, M., & Ru, L., et al. Automatic query type identification based on click through information. In Ng, H.T., Leong, M.K., & Kan, M.Y., et al. (Eds.) Information Retrieval Technology. Berlin: Springer-Heidelberg, 2006: 593–600. Retrieved on August 29, 2014 from http://link.springer.com/chapter/10.1007%2F11880592_51/. DOI: 10.1007/11880592_51.

10 Yuan, X., Dou, Z., & Zhang, L., et al. Automatic user goals identification based on anchor text and click-through data. Proceedings of the 5th Conferences of Web Information System and Application, 2008: 1–5. Xi'an. Retrieved on August 29, 2014, from http://research.microsoft.com/apps/pubs/?id=79338.

11 Lee, U., Liu, Z., & Cho, J. Automatic identification of user goals in Web search. In Lee, U., Liu, Z., & Cho, J. (Eds.) Proceedings of the 14th International Conference on World Wide Web. New York: ACM Press, 2005: 391–401. Retrieved on August 29, 2014, from http://dl.acm.org/citation.cfm?id=1060804/. DOI: 10.1145/1060745.1060804.

12 Brenes, D.J., & Gayo-Avello, D. Automatic detection of navigational queries according to behavioral characteristics. In Baumeister, J., & Atzmüller, M. (Eds.) LWA 2008: Workshop-Woche: Lernen, Wissen & Adaptivität, 2008: 1–8. Würzburg. Retrieved on August 29, 2014, from http://www.informatik.uni-trier.de/~ley/db/conf/lwa/lwa2008.html.

13 Wu, D.Y., Zhao, S.Q., & Liu, T., et al. Identification of query intents via combining multiple features. Pattern Recognition and Artificial Intelligence (in Chinese), 2012, 25(3): 500–505. Retrieved on August 29, 2014, from http://d.wanfangdata.com.cn/Periodical_mssbyrgzn201203020.aspx. DOI: 10.3969/j.issn.1003-6059.2012.03.020.

14 Maslov, M., Golovko, A., & Segalovich, I., et al. Extracting news-related queries from Web query log. In Carr, L., Roure, D.D., & Iyengar, A., et al. (Eds.) Proceedings of the 15th international Conference on World Wide Web, WWW 2006. New York: ACM Press, 2006: 931–932. Retrieved on February 1, 2015, from http://www2006.org/programme/files/xhtml/p71/pp071-maslov.html.

15 Mccreadie, R.M.C., Macdonald, C., & Ounis, I. Crowdsourcing a news query classification database. Proceedings of SIGIR'10. New York: ACM Press, 2010: 31–38. Retrieved on August 29, 2014, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.178.3236. DOI: 10.1.1.178.3236.

16 Hassan, A., Jones, R., & Diaz, F. A case study of using geographic cues to predict query news intent. Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2009: 33–41. Retrieved on August 29, 2014, from http://research.microsoft.com/en-us/um/people/hassanam/.

17 Agosti, M., Crivellari, F., & Di Nunzio, G.M. Web log analysis: A review of a decade of studies about information acquisition, inspection and interpretation of user interaction. Data Mining & Knowledge Discovery, 2012, 24(3): 663–696. Retrieved on February 1, 2015, from http://link.springer.com/article/10.1007%2Fs10618-011-0228-8. DOI: 10.1007/s10618-011-0228-8.

18 He, D.Q, & Goker, A. Detecting session boundaries from Web user logs. Proceedings of the 22nd Annual Colloquium on Information, 2000: 1–8. Cambridge. Retrieved on February 1, 2015, from http:// www.sis.pitt.edu/~daqing/docs/he00detecting.pdf.

19 Bar-Ilan, J., Zhu, Z., & Levene, M. Topic-specific analysis of search queries. Proceedings of the 2009 Workshop on Web Search Click Data. New York: ACM Press, 2009: 35–42. Retrieved on February 1, 2015, from http://dl.acm.org/citation.cfm?id=1507515. DOI: 10.1145/1507509.1507515.

20 Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20(1): 37–38. Retrieved on February 1, 2015, from http://epm.sagepub.com/content/20/1/37.full.pdf+html. DOI: 10.1177/001316446002000104.

21 Landis, J. R., & Koch, G. G. The measurement of observer agreement for categorical data. Biometrics, 1977, 33(1): 159–174. Retrieved on August 29, 2014, from http://www.jstor.org/stable/2529310. DOI: 10.2307/2529310.

22 Herrera, M.R., de Moura, E.S., & Cristo, M., et al. Exploring features for the automatic identification of user goals in Web search. Information Processing and Management, 2010, 46: 131–142. Retrieved on February 1, 2015, from http://www.sciencedirect.com/science/article/pii/S0306457309001058. DOI: 10.1016/j.ipm.2009.09.003.

23 Dalianis, H., Xing, H., & Zhang, X. Creating a reusable English-Chinese parallel corpus for bilingual dictionary construction. Proceedings of the International Conference on Language Resources and Evaluation, 2010: 1–4. Valletta, Malta. Retrieved on February 1, 2015, from http://www.researchgate.net/publication/220746929_Creating_a_Reusable_English-Chinese_Parallel_Corpus_for_Bilingual_Dictionary_Construction.

24 Zhao, Q., Liu, T.Y., & Bhowmick, S.S., et al. Event detection from evolution of click-through data. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06). New York: ACM Press, 2006: 484–493. Retrieved on August 29, 2014, from http://dl.acm.org/citation.cfm?id=1150456. DOI: 10.1145/1150402.1150456.

25 Sun, A., & Hu, M. Query-guided event detection from news and blog streams. IEEE Transactions on Systems, Man, and Cybernetics (Part A) 2011, 41(5): 834–839. Retrieved on August 29, 2014, from http://www.bibsonomy.org/bibtex/22294883c3052c484ee505ae863991206/dblp.

26 Claypool, M., Brown, D., & Le, P., et al. Inferring user interest. Proceedings of IEEE Internet Computing. 2001, 5(6): 32–39. Los Alamitos: IEEE Computer Society, Retrieved on August 29, 2014, from https://www.zotero.org/silverasm/items/itemKey/B5T8Q36Z.

27 Gaugazl, J., Siehndel, P., & Demartini, G., et al. Predicting the future impact of news events. Proceedings of the 34th European Conference on Advances in Information Retrieval. Berlin: ACM Press, 2012: 50–62. Retrieved on August 29, 2014, from http://dl.acm.org/citation.cfm?id=2260648. DOI: 10.1007/978-3-642-28997-2_5.

28 Ruocco, M., & Ramampiaro, H. Exploratory analysis on heterogeneous tag-point patterns for ranking and extracting hot-spot related tags. Proceedings of the SIGSPATIAL. New York: ACM Press, 2012: 16–23. Retrieved on August 29, 2014, from http://dl.acm.org/citation.cfm?id=2442802. DOI: 10.1145/2442796.2442802.

29 Brenes, D.J., Gayo-Avello, D., & Perez-Gonzalez, K. Survey and evaluation of query intent detecting methods. Workshop on Web Search Click Data (WSCD'09). New York: ACM Press, 2009: 1–7. Retrieved on August 29, 2014, from http://dl.acm.org/citation.cfm?id=1507510. DOI: 10.1145/1507509.1507510.

# Submission Guidelines

## ◆ Aims

*Chinese journal of Library and Information Science* (*CJLIS*), being sponsored by the Chinese Academy of Sciences (CAS) and published quarterly by the National Science Library of CAS, is a scholarly journal in the field of library and information science (LIS). Its aim is to provide an international communication link between researchers, educators, administrators, and information professionals.

With the publication of the research results both from China and from other foreign countries, the Journal *CJLIS* strikes a balance between theory and practice. With its goal to provide an open forum for Chinese and international scholars in this field to exchange their research results, *CJLIS* also offers new possibilities in the advancement of Chinese library operations. The *CJLIS* tries to establish a platform for LIS students, researchers and library staff all over the world to engage in intellectual dialog and also to improve library services so as to promote even more quickened and substantial development of LIS in China.

## ◆ Scope

Striving toward academic excellence, innovation, and practicality, the *CJLIS* mainly includes research papers both on the theoretical as well as on the practical fronts in all aspects of the field. More specifically, it includes but not limited to informatics, library management, information technology application, knowledge organization system, knowledge management, archives, permanent preservation of library resources, LIS education, and so on.

## ◆ Refereeing Process

Articles and papers covering the topics or themes mentioned above will be refereed through a double-blind peer review process.

## ◆ Editorial Advisory Board

The Editorial Board is composed of the nationally and internationally well-known scholars and researchers in the LIS field and the high quality of this Journal is thus reasonably assured.

## ◆ Manuscripts Categories

As the first English-language academic journal on LIS published in Mainland China, the *CJLIS* will take a proactive attitude to trace and report the prevailing hot issues in the field around the globe as well as the more serious scholarly communications. As such, the submitted manuscripts are classified into constant categories and unfixed categories. In the former category, research papers, library practice and progress reports are the essential components. In the latter, book reviews, biographical sketches, anecdotes, reminiscence of prominent librarians and brief communications will appear occasionally.

***Research papers*** represent original research work or a comprehensive and in-depth analysis of a topic. More than 3,000 words are considered as a proper length for such manuscripts, with a structured abstract ca. 200 words.

***Library practice*** covers the latest development and application in any segment of library field work and information service. The length of the manuscript is preferred to be more than 3,000 words, with a structured abstract ca. 200 words.

***Progress reports*** reflect the projects result or research progress on the key topics of the library and information science. Submissions of articles to this section are expected to be comprehensive and analytical, which may deepen the understanding of the discussed issue and stimulate further researches on the topics, or give a new perspective on future technological applications. The manuscript length should be within 5,000 words, with a structured abstract ca. 200 words.

## ◆ Manuscripts Requirements

All papers can be submitted either in English or in Chinese (or both) with a double-line space. For the assurance that all the materials of the to-be-submitted are included, please check the following:

**Title.** Please give a brief biographical introduction to all contributing authors and their research background on a separated paper. For a better organization of the paper, please use the headings and subheadings.

**Authors and affiliations.** Please do not forget to write down the mailing address of each and every article contributors.

**References.** Be sure all the references used should be cited properly in both in-text and in bibliography. Particular attention should be paid to the proceedings. Do not forget adding the name(s) of editors of the compilation, as well as the name of the publishers. For the detailed information, please request a copy of **Reference Citation Format**.

## ◆ Copyright

All submitted papers normally should not have been previously published nor be currently under consideration for publication elsewhere. For all the materials translated or obtained from other published resources, they should be properly acknowledged. All copyright problems should be cleared without any legal entanglements prior to the publication.

## ◆ Notes for Intending Submissions

A guide for authors and other relevant information, including submitting papers online, is available at the website of the Editorial Office of the *CJLIS* (http://www.chinalibraries.net). For any questions, you can e-mail the Office or directly to:

Prof. ZHANG Xiaolin
Editor-in-Chief of *CJLIS*
The *CJLIS* Editorial Office
National Science Library, Chinese Academy of Sciences
No.33 Beisihuan Xilu, Zhongguancun, Haidian District,
Beijing 100190, P.R. China
Tel: 86-10-82624454 or 86-10-82626611 ext. 6628
Fax: 86-10-82624454
E-mail: chinalibraries@mail.las.ac.cn
Website: http://www.chinalibraries.net

## ◆ Subscription

For single-copy subscription in China: RMB ¥ 200/Issue. For subscription outside of China, US $ 199/Volume yearly (including air shipping)

# Chinese Journal of Library and Information Science (Quarterly)

Volume 7
Number 4
(December 2014)


## Contents

### Research Papers




ISSN 1674 – 3393
CN 11-5670/G2