SOGOU-2012-CRAWL: A Crawl of Search Results in the Sogou 2012 Chinese Query Log

Stewart Whiting & Joemon M. Jose
School of Computing Science,
University of Glasgow,
Scotland, UK.
{stewh,jj}@dcs.gla.ac.uk

Omar Alonso
Microsoft Corp.
Mountain View,
California, USA.
omalonso@microsoft.com

ABSTRACT

In 2012, Sogou, a major Chinese web search engine released a large-scale query log containing 43.5M user interactions, including submitted queries and clicked web page search results. This query log offers a deep sample of queries over a two day period from 30th December 2011 to 1st January 2012. In August 2013, we identified 1.4M predominantly Chinese language unique search result URLs that were clicked at least three times in this query log. We crawled the HTML content of these URLs to construct the supplementary SOGOU-2012-CRAWL dataset, which we release in this work. A real large-scale query log with accompanying crawl such as this offers several opportunities for reproducible information retrieval (IR) research, including query classification, intent modelling and indexing strategy. In this paper we first detail the query log and crawl dataset construction and characteristics. Following this, to demonstrate potential applications we use the crawl to indicatively analyse various time-based patterns in web content and search behaviour. In particular, we study the distribution of language-independent date expressions in the crawled web content. Based on this, we propose a simple approach for modelling the past/present/future temporal intent of queries based on the date the query was submitted by the user, and the dates appearing in the clicked search results. We observe several prominent temporal patterns which may lead to novel time-aware IR approaches.

1. INTRODUCTION

A great deal of research in information retrieval (IR) is based on analysing the past behaviour of real IR system users. Such user interaction is captured in a *query log* which comprises a rich time-stamped record of past user actions such as query input, result clicks and result page navigation. From the lowest-level interactions, higher-level task-oriented aspects such as information seeking sessions, topic focus, intent and implicit relevance can be observed, studied and modelled.

Unfortunately, despite their research value, few query logs are available for researchers outside industrial labs due to privacy concerns and commercial sensitivity. Indeed, the most recent large-scale English-language query logs are AOL [9] and MSN 2006 [3]. More recent query logs are available for learning to rank tasks - however their queries and result domains/URLs have been mapped to arbitrary tokens to ensure privacy, thereby making content-based analysis impossible.

In recent years, Sogou[1], a major Chinese portal and web search engine, has made several large-scale Chinese web search engine query logs openly available for research. While this means that some analysis inevitably requires Chinese language skills, many findings are likely to be generalisable to IR systems in other languages since they are based on language-independent behavioural, temporal and textual statistical distributions.

To the best of our knowledge, there is no sufficiently large nor recent query log with accompanying result crawl openly available. A query log without result crawl impedes deeper content analysis of search results for tasks such as query classification, intent modelling and indexing strategies. Hence, in this work we present a supplementary result crawl dataset to augment the Sogou 2012 query log with the full HTML content of clicked result URLs – thereby offering many opportunities for novel research.

The Sogou 2012 query log contains 43.5M low-level interactions (e.g., queries submitted, results clicked with their ranking, and result page inspection depth) from 0:05am 30th December 2011 to 4:58pm 1st January 2012. Accordingly, the query log offers considerable depth into search activity over a relatively short period, as opposed to a small sample of search activity over a longer period as is more common in previous datasets. The query log only contains search results clicked - so analysis of non-clicked results cannot be performed. The dataset we prepared and present in this paper, named SOGOU-2012-CRAWL, is a crawl of 1.4M search results clicked in this Sogou query log. We detail the construction and release of this dataset in Section 2.

To indicatively demonstrate the potential of SOGOU-2012-CRAWL for IR research, we study some of its language-independent content and query-biased temporal dimensions. In particular, we consider the following questions: (i) to what extent do web pages contain dates? (ii) how are the dates distributed in those web pages? (iii) can the temporal intent of a query (or similarly, its *topic*) be determined based on the dates contained in clicked, and therefore assumed implicitly relevant web pages?

2. SOGOU-2012-CRAWL DATASET

In this section we detail the construction of the SOGOU-2012-CRAWL dataset. We have made this supplementary crawl dataset available to download as a 15Gb archive of HTML page snap-

[1]http://www.sogou.com

SIGIR '16, July 17-21, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2911451.2914668

shots[2]. The Sogou 2012 query log itself must be downloaded directly from the Sogou labs website[3].

Because of the long-tail popularity distribution of result clicks, there is a very large number of results clicked only once or twice in the query log. From a practical perspective, since our crawler was based geographically in Europe we found that network speed, reliability and time outs to servers located inside China posed several crawling technical difficulties. To alleviate these issues by reducing the number of web pages we needed to download, we only crawled search result URLs that were clicked three or more times by users (for any given search query). Accordingly, we identified 1.41M unique search results for crawling.

In August 2013, we successfully crawled 1.22M of the URLs (87% success rate) to form the SOGOU-2012-CRAWL dataset. Although not always the case, we have to assume the URLs have not changed since the the result click as the vast majority of the documents are not in the Internet Archive[4] for the result click date. This assumption will undoubtedly be invalid for news sites which change daily – so temporally sensitive analysis (such as that we perform in the following sections) must be conducted with care.

Several data cleaning stages were employed to prepare the final crawl dataset for release. Many result URLs were unable to be crawled because the server was uncontactable (e.g., server offline), an error page was returned (e.g., 401, 404, 500 HTTP errors), or had a robots.txt entry that excluded crawling. In these cases, the result URL still appear in the crawl dataset but with content of '[notavailable]' to signify a crawling issue.

Several character set encodings are commonly used in China – in particular UTF8, GBK and GB2312. In many cases, the character set reported by HTTP headers or HTML head tags, and the actual content encoding were mismatched. We used various techniques to detect this, and encode the final crawl dataset consistently in UTF8.

3. CRAWL TEMPORAL ANALYSIS

To demonstrate the potential value of the SOGOU-2012-CRAWL dataset, we perform preliminary analysis of its temporal characteristics. First we extract and analyse the dates present in result pages in the crawl as a whole – that is, without considering the related query times. More interestingly, in Section 4 we consider the temporal information provided by the crawl and the query log together. With that in mind, we further study how the dates present in clicked result URLs correspond with the real-time of the search queries, in order to model query-biased temporal search intents.

3.1 Related Work

In this section, we present existing work related to temporal analysis of information seeking and web content in IR. *When* a query is posed to a system provides an implicit temporal clue, since it offers a signal of the user's current interest in a particular query topic. Likewise, temporal clues such as dates and times contained in relevant content also provide an anchor in time for that relevance. Past work has mostly examined these elements independently.

The popularity of many queries is subject to periodic temporal trends [1], as well as event-driven temporal activity arising from one-off events. Explicit temporal expressions are included in up to 1.5-2% of web queries [8], such as "`world cup 2012`" and "`nyc weekend events`". [10] consider these temporal expressions in a relevance model. Further, [7] do the same for query topics which implicitly express a temporal expression (e.g., "`brazil world cup`"). [2] integrate temporal expressions in documents into a time-aware probabilistic retrieval model.

[6] identify and classify temporal information needs based on the relevant document timestamp distribution to improve retrieval. [5] studied how 150M web pages changed over 11 weeks. They found longer documents tended to have more frequent and extensive changes made to them, and for web pages that change regularly, they do so consistently over time. Acknowledging web page content is often dynamic, [4] incorporate content dynamics into a relevance model to improve navigational query effectiveness.

[2]http://www.stewh.com/sogou-2012-crawl

[3]http://www.sogou.com/labs - note that this site is available in English and Chinese language versions.

[4]http://www.archive.org

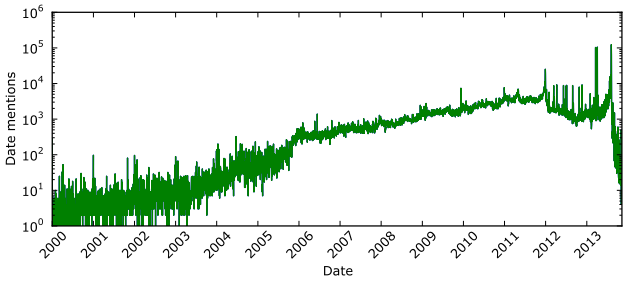

Figure 1: Frequency distribution (from 2000-01-01 to 2013-12-31) of dates contained in 1.22M search result pages crawled from Sogou 2012 result click URLs. Plotted with a log_{10} y-axis frequency scale.

3.2 Extracting Temporal Expressions

For temporal analysis, we must first extract temporal expressions (in this case, date expressions) appearing in the crawled result pages. Dates can be expressed in many cultural short- and long-forms. In written Chinese language, dates are typically in big-Endian Chinese year/month/day date formats, e.g. `YYYY-MM-DD`, `YYYY.MM.DD` and `YYYY年MM月DD日`.

From the complete crawl dataset, we extracted 5.41M dates between 2000-01-01 to 2014-12-31, inclusive, from a total of 313,211 web pages. All temporal analysis presented in this paper is based on these extracted dates – and in the next section, the queries that were used to find the pages containing those dates. For the purpose of this work, we relied on simple temporal expression extraction based on regular expressions. Future work will employ full multi-lingual and diverse temporal expression tagging, such as that provided by HeidelTime [11], to improve coverage and accuracy.

Overall, 25.7% of the crawled result web pages contain at least one date, with an average of 17.27 ($\pm$59.7) and median of 5 dates per page. Note the wide dispersion in the number of dates in each web page – the majority of documents contain a few dates, while a smaller amount have a very large number of dates. Web pages with a single date are likely to be timestamped content, such as a news article or blog post. Likewise, documents with several dates may be continually revised information, such as blog posts with comments, or similarly threaded discussion boards.

The frequency distribution of all the extracted dates over time is presented in Figure 1. Evident is the rising frequency of dates up to the end of 2011 – the period of time covered by the query log. Following this, the frequency of dates declines quickly, although some of this volume may be attributed to the delayed crawling. Occasional spikes likely represent major upcoming predictable events being discussed (e.g., major sporting events, public holidays etc.).

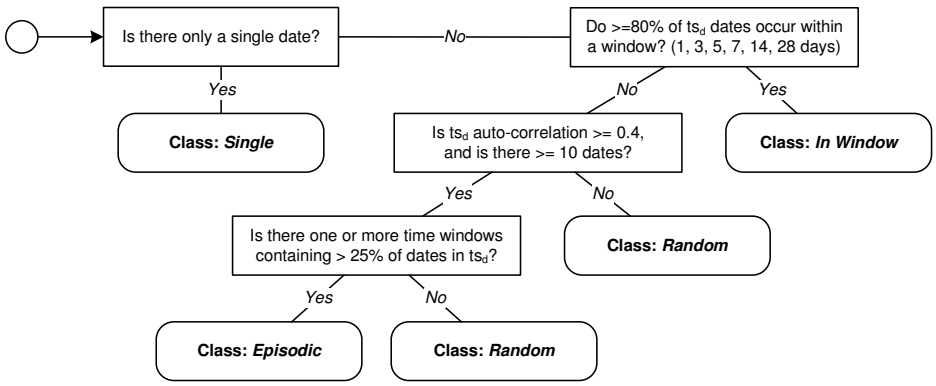

Figure 2: Heuristic-based decision tree classifier for categorising crawled search result pages into temporal date distributions.

3.3 Document Dates Distribution Classes

In the next step of temporal analysis we characterise the distribution of dates occurring in each *individual* web page in the crawl. To that end, in this section we propose a classifier for 4 common document date distribution classes we have observed: *Single*, *In Window*, *Episodic* or *Random*. Each of these classes is conceptually defined as follows:

Single - web page contains a single date expression.

In Window - web page contains multiple date references, with $\geqslant$ 80% of dates occurring within an n day period ($n = 1, 3, 5, 7, 14, 28$ days). An 80% (rather than 100%) threshold permits more robust tolerance to outlying dates (for instance, a '*today*' date on a page containing mainly past timestamped discussion).

Episodic - web page has significant trends in date distribution, with 1 or more defined bursts.

Random - web page contains multiple dates, but with no immediate apparent temporal pattern.

3.4 Classifier Implementation

We use a classifier to label each web page containing dates with one of the above temporal classes. To that end, we first define ts_d as the time series of dates present in the document d in the our crawl. ts_d is modelled as vector of length 5,475 (i.e., for the 365 days $\times 15$ years, between 2000-01-01 and 2014-12-31). Consequently, the ith index of the ts_d vector refers to number of dates in the document on the day 2000-01-01, plus i days.

We use a decision tree (shown in Figure 2) incorporating various heuristics to determine the class of each page, based entirely on features extracted from ts_d. Auto-correlation is the cross-correlation of the signal with a lagged version of itself (1 day in this case). It is effective for measuring temporal uniformity, and co-ordinated bursts of dates [6]. Single and In Window (1 day window) classes are in essence the same, except that the In Window class applies to pages which have multiple instances of the same date.

3.5 Results and Discussion

In Table 1, we present the results of classifying 313,211 web pages containing at least one date. Almost 16% of these pages have a discernible pattern comprising multiple dates contained within them. Single and In Window classes account for almost all documents, with results presented in Table 2 showing a windows of 1-3 days are most common for those pages. As such, we find that for the almost 50% of search result web pages containing non-randomly distributed dates, the majority contain either a single date, or several dates contained in a short period of days.

Table 1: Date distribution classification results for the 313,211 search result pages containing dates.

Class	Instances	%
Single	98,177	31.1%
In Window	47,831	15.3%
Episodic	3,317	1.1%
Random	163,886	52.3%

Table 2: Classification window sizes for the 47,831 search result pages containing multiple dates in a window of 28 or less days.

Window Size	Instances	%
1 day	25,249	52.8%
3 days	5,217	10.9%
5 days	2,276	4.8%
7 days	2,884	6.0%
14 days	6,733	14.1%
28 days	5,472	11.4%

4. QUERY-BIASED TEMPORAL INTENTS

In many cases, users search for topics or similiarly, events, anchored in the past, present or future with respect to the present time [6, 7]. We seek to identify these intents by examining the dates present in the search results of clicked (and therefore, assumed relevant) search results. Rather than classifying the distribution of dates present in a single document, in this section we study the distribution of dates contained in the assumed relevant documents for a topic (that is, clicked search results) – with respect to the real-time of the query (which we refer to as q_τ).

The Sogou 2012 query log contains 42.17M search result clicks in total. 3.11M (7.4%) of these are for documents which were classified as Single/In Window/Episodic in the previous section (i.e., have a non-random date distribution)[5].

Considering each query q and the day it was used as unique (i.e., q_i will be counted twice if it appears on two separate days), there are 10.1M queries in the dataset. 1.1M queries (10.9%) have one or more clicked search result which contains at least one date.

We define the *query-biased temporal intent* (or, $q_{intent,\tau}$) in terms of the average difference (Δ, in days) between q_τ and dates appearing in clicked documents (weighted by their click proportion).

[5]Recall, we only crawled search result URLs with 3 or more clicks, so the long-tail may substantially increase this percentage.

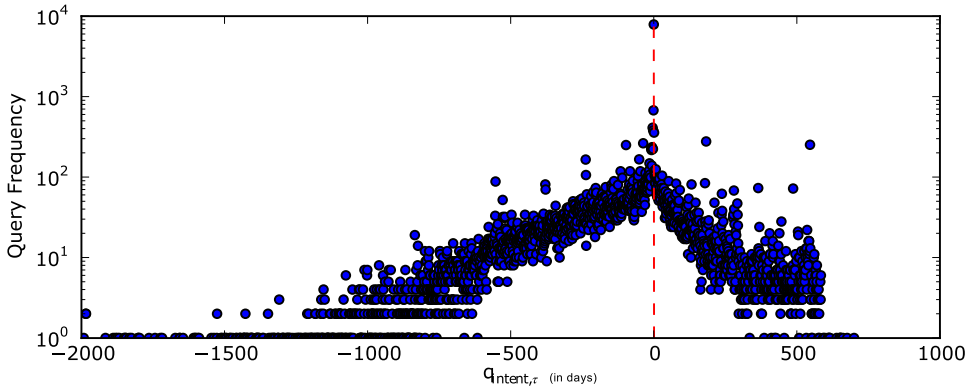

Figure 3: $q_{intent,\tau}$ (that is, temporal intent in terms of days past/present/future from query time) distribution of 1.1M queries in Sogou 2012. Data points less than -2000 days removed to improve graph clarity. Vertical marker at $x = 0$ denotes '*today*' temporal intents.

This means $q_{intent,\tau}$ may be negative (i.e., an intent in the *past*), positive (i.e., an intent in the *future*), or zero (i.e., an intent *now/today*). Note that $q_{intent,\tau}$ may also be undefined in the case that there are no dates present in any of its clicked documents. $q_{intent,\tau}$ is formalised as follows. For each query q_i occurring at time q_t (we specify time by day, e.g., 2011-12-30), we compute $q_{intent,\tau}$ as:

$$q_{intent,\tau} = \frac{\sum_{d_i \in D_q} \sum_{\tau_i \in dates(d_i)} \Delta(\tau_i, q_\tau)}{\sum_{d_i \in D_q} |dates(d_i)|} \quad (1)$$

Where D_q is the set of all document results d_i (including duplicates) clicked for q_i, i.e., $D_q = \{d_1, d_2 \ldots d_i\}$. $dates(d_i)$ is the set of dates τ_i (including duplicates) present in d_i, i.e., $dates(d_i) = \{\tau_1, \tau_2 \ldots \tau_i\}$. Finally, $\Delta(x, y)$ is the total days between two dates.

4.1 Results and Discussion

In Figure 3, we show the $q_{intent,\tau}$ distribution of 1.1M queries with clicked search results which contain dates in Sogou 2012. A distinctive distribution of past, present and future query-biased temporal intents is evident. Most notably, there is a search tendency centred around '*today*' dates (that is, the time of the query interaction, i.e., 30th December 2011 to 1st January 2012), with very recent past and upcoming future dates also common. The steady decline in queries with far mid- to long-term past and future intents is also apparent. There are notably more search intents for past topics, with the majority of intents going back the past 1,000 days. In comparison, the majority of future intents goes forward 500 days. Outliers are likely notable past/future events, for example the 2012 London Olympics. These event topics attract lots of different queries and have many dates mentioned in search results referring to published event schedules.

5. DISCUSSION AND CONCLUSION

In this paper we present a major new dataset – a supplementary crawl of the web search results clicked in the open, large-scale and predominantly Chinese language Sogou 2012 query log. We name this dataset SOGOU-2012-CRAWL, and release it to the research community for future work.

While there are several query logs openly available for reproducible research, to the best of our knowledge, there is no recent large-scale query log accompanied by the content of clicked search results. As a result, this dataset opens up several new opportunities for information retrieval research challenges such as query classification, intent modelling and indexing strategy.

To characterise the crawl, and demonstrate its value for language-independent IR research, we indicatively analysed various time-based patterns in web content and search behaviour. We first studied the distribution of language-independent date expressions in the crawled web content. Further, we propose a simple approach for modelling the past/present/future temporal intent of queries based on the date the query was submitted by the user, and the dates appearing in the clicked search result pages. Based on this model we observe several notable query-biased temporal intent patterns. In particular, we note that users' temporal interests are mainly concentrated on information related to *now*. However, we find there are a large number of interests which extend into the future, and likewise, up to 1,000 days into the past (shown in Figure 3). This asymmetric double long-tail distribution may offer opportunities for detecting and supporting time-aware information needs, so we leave further analysis and modelling to future work.

6. REFERENCES

[1] S. M. Beitzel, E. C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman. Temporal analysis of a very large topically categorized web query log. *J. Am. Soc. Inf. Sci. Technol.*, 58(2):166–178, 2007.

[2] K. Berberich, S. Bedathur, O. Alonso, and G. Weikum. A language modeling approach for temporal information needs. ECIR'2010, pages 13–25, Berlin, Heidelberg, 2010. Springer-Verlag.

[3] N. Craswell, R. Jones, G. Dupret, and E. Viegas. Wscd '09: Proceedings of the 2009 workshop on web search click data. New York, NY, USA, 2009. ACM.

[4] J. L. Elsas and S. T. Dumais. Leveraging temporal dynamics of document content in relevance ranking. WSDM '10, pages 1–10, New York, NY, USA, 2010. ACM.

[5] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. WWW '03, pages 669–678, New York, NY, USA, 2003. ACM.

[6] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3):14–es, 2007.

[7] N. Kanhabua and K. Nørvåg. Determining time of queries for re-ranking search results. ECDL'10, pages 261–272, Berlin, Heidelberg, 2010. Springer-Verlag.

[8] S. Nunes, C. Ribeiro, and G. David. Use of temporal expressions in web search. ECIR'08, pages 580–584, Berlin, Heidelberg, 2008.

[9] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. InfoScale '06, New York, NY, USA, 2006. ACM.

[10] J. Strötgen, O. Alonso, and M. Gertz. Identification of top relevant temporal expressions in documents. TempWeb '12, pages 33–40, New York, NY, USA, 2012. ACM.

[11] J. Strötgen and M. Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. SemEval '10, pages 321–324, Stroudsburg, PA, USA, 2010. ACL.