Detecting Seasonal Queries Using Time Series and Content Features

Behrooz Mansouri
School of Electrical and Computer Engineering, University of Tehran
Iran
b.mansouri@ut.ac.ir

Mohammad Sadegh Zahedi
School of Electrical and Computer Engineering, University of Tehran
Iran
s.zahedi@ut.ac.ir

Maseud Rahgozar
Database Research Group, Control and Intelligent Processing Center of Excellence, School of Electrical and Computer Engineering, University of Tehran
Iran
rahgozar@ut.ac.ir

Ricardo Campos
Polytechnic Institute of Tomar
LIAAD INESC TEC
Portugal
ricardo.campos@ipt.pt

ABSTRACT

Many user information needs are strongly influenced by time. Some of these intents are expressed by users in queries issued indistinctively over time. Others follow a seasonal pattern. Examples of the latter are the queries "*Golden Globe Award*", "*September 11th*" or "*Halloween*", which refer to seasonal events that occur or have occurred at a specific occasion and for which, people often search in a planned and cyclic manner. Understanding this seasonal behavior, may help search engines to provide better ranking approaches and to respond with temporally relevant results leading into user's satisfaction. Detecting the diverse types of seasonal queries is therefore a key step for any search engine looking to present accurate results. In this paper, we categorize web search queries by their seasonality into 4 different categories: Non-Seasonal (NS, e.g., "*Secure passwords*"), Seasonal-related to ongoing events (SOE, "*Golden Globe Award*"), Seasonal-related to historical events (SHE, e.g., "*September 11th*") and Seasonal-related to special days and traditions (SSD, e.g., "*Halloween*"). To classify a given query we extract both time series (using the document publish date) and content features from its relevant documents. A Random Forest classifier is then used to classify web queries by their seasonality. Our experimental results show that they can be categorized with high accuracy.

KEYWORDS

Temporal IR; Temporal Query Classification; Seasonal Queries;

1 INTRODUCTION

The importance of time cannot be neglected on web search. The temporal intent of the searcher adds an important dimension to the relevance judgments of web queries. However, lack of understanding their temporal requirements, increases the difficulty in clearly understand the real intention behind the user's query. Identifying the query temporal nature offers search engines the chance to provide better ranking approaches leading into user's satisfaction. For example, a search engine retrieving results for the query "*Halloween*" will mostly return recent pages related to this event during peak times (a likely indication of the approaching of the event), while Wikipedia-like pages during non-peak times. In this work, we are particularly concerned in identifying seasonal queries, a sub-type of temporal queries. Seasonal queries are issued periodically usually triggered by expected events occurring at specific planned and repeated occasions. During the occurrence of the event or on the days before or after it, an increase in the number of published documents and queries that concern the event can be observed. Figure 1 shows an example of this user behavior for the queries "*Golden Globe Award*" (Seasonal-related to ongoing event), "*September 11th*" (Seasonal-related to historical events) and "*Halloween*" (Seasonal-related to special days and traditions) from January 2004 (which marks the beginning of the Google Trends feature) to April 2017. The figure portraits a time series built by query frequency volume. A large number of spikes - mostly corresponding to the occurrence or celebration of the event - can be observed for each query, thus proving its seasonality.

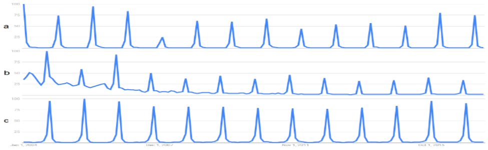

Figure 1: Query frequency pattern examples from Google Trends from January 2004 to April 2017. (a: "Golden Globe Award", b: "September 11th", c: "Halloween").

The genesis of detecting seasonality, dates back to 2010, when Zhang et al. [14] proposed to detect recurrent queries of related-events occurring at predictable time intervals, by using a machine learning classifier which is built on top of query logs, search sessions, clicks and time series features. A similar work was proposed by Shokouhi [13] who used seasonality of query volume time series to detect seasonal queries. In their work, they used time series decomposition techniques to measure seasonality of queries. In addition to the methods above,

ICTIR '17, October 1–4, 2017, Amsterdam, Netherlands
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4490-6/17/10...$15.00
https://doi.org/10.1145/3121050.3121100

Kulkarni et al. [16] analyzed a query log over the course of 10 weeks to explore how query's intent change over time. There has also been substantial work involving the dynamics and the classification of time-sensitive queries. The temporal dynamics of web queries have been commonly studied by building time series for queries based on their past frequency at uniform intervals and extracting time series features [5,8,13]. An interesting tutorial on this topic has been given by Radinsky et al. [12]. Other than frequency volume, previous researches used click log, query reformulation and relevant documents to better understand user temporal intent [1, 3, 11, 14]. In particular, Jones and Diaz [8], introduce a model to measure the distribution of documents retrieved in response to a query over the time domain in order to create a temporal profile for a query. They introduced three temporal classes of queries: atemporal, temporally ambiguous and temporally unambiguous. Campos et al. [1] also propose to classify queries into one of these three categories using information extracted from web snippets. Metzler et al. [11] in turn, used query logs to investigate implicitly year qualified queries. The work by Gupta and Berberich [5] describes a taxonomy of temporal classes at different granularities. Ghoreishi and Aixin [3] and Kanhabua et al. [9] studied event-related queries within Temporalia task of NTCIR [7] which considers 4 classes: atemporal, past, present, and future. A fully detailed description on Temporal IR applications can be found in the survey of Campos et al. [2]. In the next section, we present our classification taxonomy. The remainder of this paper is organized as follows. Section 3 describes the features used for classification. Section 4 outlines our experiment results. Finally, Section 5 provides some conclusions.

2 SEASONAL QUERIES CLASSIFICATION

Despite previous attempts to tackle the problem of identifying seasonal queries, no one so far, has considered to use content features as a means to understand the reasons for seasonality. Different approaches have been presented for this purpose [13 - 14], however mostly focused on identifying seasonal queries based on time series and query reformulation data. In this work, we plan to use relevant published documents as a way to capture seasonality. In particular, we use time series and content features to capture valuable information. This may be understood as an important contribution to the community. As an additional contribution, we present a new taxonomy which distinguishes between the different types of seasonal queries: Non-Seasonal (NS), Seasonal-related to ongoing event (SOE), Seasonal-related to historical events (SHE) and Seasonal-related to special days and traditions (SSD). In particular, NS refer to those types of queries that do not show any seasonal spike in their related time-series (e.g., “passwords”); SOE concern events that in each episode a new story - which is different from previous ones - happens (e.g. “US Presidential Elections”, “Olympics”); SHE shows periodic spikes because of an old, usually historical event, for which users are tempted to search whenever the celebration date approaches (e.g. “September attacks”, “Adolf Hitler Death”, “Iranian Revolution”). Finally, SSD concern special days and traditions (e.g. “Halloween”, “Thanksgiving Day”).

Detecting the different types of seasonal queries can be very useful for search engines aiming to adapt their results depending on the type of seasonal query detected. Our assumption, which we will confirm by query log analysis, is that different queries, despite having the same time series shape (recall Figure 1), may present different requirements in terms of the results to be returned to the user. For example, seasonal queries such as “Halloween” may require more recent pages during peak times, whereas an historical seasonal query such as “September 11th”, which is often issued by users each year, demands, instead, more Wikipedia-like pages. To validate our assumption, we conduct a user survey on two-year query log of a Persian commercial search engine. We studied user's behavior towards seasonal queries under peak and non-peak times, where peak time is defined as the time that goes from a week before and a week after the occurrence of the seasonal event, and non-peak time is defined as the rest of the time that do not fit within the previous interval. For this purpose, we selected 150 seasonal queries with a two-year query log frequency higher than 100. These queries concern well-known seasonal events that took place on repeated occasions (e.g., “Halloween”) during this two-year time period. All the queries were then manually classified into each one the three seasonal classes by 4 professional editors. An inter-rater reliability analysis using the Fleiss Kappa statistics was performed to determine consistency among the editors. Overall, the annotators obtained about 0.88 of agreement level, which represents a high agreement between editors. 41 queries were labeled SHE (Seasonal-related to historical events), 50 SOE (Seasonal-related to ongoing event) and 59 SSD (Seasonal-related to special days and traditions). For each query, we considered the top-200 clicked pages (100 pages from peak time and 100 from non-peak time) totalizing 30.000 web pages. We then asked one student to look at the content of each web page and to manually classify them with regards to recency, oldness and Wikipedia-like page (a type of page that is usually retrieved for seasonal events). In particular, each web page is classified into: (1) Recent Pages; which provide information about the most recent episode of the event, (2) Wikipedia-Like Pages; which gives general information about the event and (3) Old Pages; which concerns the old episodes of the event. Table 1 summarizes the result of our study, by showing the percentage of clicked pages per page categories during peak and non-peak times for each seasonal query class.

Table 1: Percentage of clicked pages per seasonal queries during peak and non-peak times.

Seasonal Query Class	Recent Pages		Wikipedia-Like Pages		Old Pages	
	Peak	Non-Peak	Peak	Non-Peak	Peak	Non-Peak
SOE	**92.1%**	51.4%	2.5%	7.1%	5.4%	41.5%
SHE	54.7%	7.3%	44.9%	**90.6%**	0.4%	2.1%
SSD	**94.3%**	4.9%	4.1%	91.3%	1.6%	3.8%

Based on our experiment, we were able to confirm that, despite having the same time series shape, different seasonal queries may require a different type of results. Thus, detecting the different type of seasonal queries may reveal an important feature for any search engine looking to provide better ranking approaches. Our experiments on query logs show that the results to be retrieved should differ during peak and non-peak times for each seasonal category class. Observing the results for SOE queries one can conclude that during peak times, users prefer most recent web pages thus retrieving more fresh documents seems to be the best choice. This contrasts with non-peak periods, for which temporal diversity is suggested. Likewise, considering peak times on SHE queries, users are mostly interested in getting to know about recent commemorative and memorial gatherings. However, in contrast to SOE queries, a notable amount of users were also interested in Wikipedia-like pages. On non-peak times, Wikipedia-like pages were also dominantly clicked compared to other type of web pages. Finally, SSD queries were mostly favored with recent pages during peak times, while Wikipedia-like pages were preferred on non-peak times.

3 OUR APPROACH

To detect the different types of seasonal queries, we expand the queries with two types of features: (i) time-series; (ii) content features.

3.1 Time Series Features

A time series is a sequence of values of a particular measure taken at regularly spaced intervals over time. In the context of web search, a time series can be constructed for a query based on the queries past frequency or generated on top of the retrieved documents published time. In this work, we chose the latter. Thus, instead of resting on query log features, we follow a metadata-based approach which rests on top the documents published time. Here we introduce our 7-time series features: (1) **Autocorrelation** indicates how well a time series is similar to a time-shifted copy of itself. We used lag-1 autocorrelation of a time series which is the correlation of each value with the immediately preceding observation. Time series of queries with strong inter-day dependency have higher autocorrelation value [10]. Autocorrelation of time series T with lag=1 can be calculated as follows ($\bar{t}$ is the mean value of time series):

$$Autocorrelation(T) = \frac{\sum_{i=1}^{N-1}(t_i - \bar{t})(t_{i+1} - \bar{t})}{\sum_{i=1}^{N}(t_i - \bar{t})} \quad (1)$$

(2) **Seasonality** represents the cosine similarity between time series and its seasonal component. Different decomposition approaches can be applied to time series in order to analyze its seasonal components. In this work, we use Holt-Winters decomposition technique [4]. After decomposing the time series, we remove its trend component from the time series as it just shows the overall trend of a query and then calculate the cosine similarity between the seasonal component and the remaining components. Considering S as seasonal component of time series and $\hat{T}$as time series with removed trend component seasonality of time series T is:

$$Seasonality(T) = \frac{S \,.\, \hat{T}}{\|S\| \,.\, \|\hat{T}\|} \quad (2)$$

(3) **Kurtosis** calculates how much of the probability distribution is contained in the peaks and how much in the low-probability regions [8] and is calculated as the ratio of the fourth moment and variance squared. (4) **Randomness Test** is used to analyze the distribution of a set of data to see if it is random. We calculate *p*-value of Mann-Kendall rank test [11] and use it as a feature of randomness. (5) **SSE** (Sum of Squared Errors) of a prediction model can show how the time series is unplanned at a given point. We estimate predicted values using Holt-Winters [4] approach. (6) **Modality** in time series show number of detected modes. Seasonal queries should have multi-modal time series. In our work, we used Dip test [6] to calculate number of modes. (7) **Mean** value of time series.

3.2 Content Features

In addition to time series features (over a metadata-based approach), we also consider six content features: (1) **Content clarity:** shows how specific a query is, and it is measured by calculating the KL-divergence between the collection language model and the relevant documents language model. A higher KL-divergence value indicates that the query is clear and that its related documents concern a more specific topic. (In this work, we used unigram language model). (2) **Year expressions:** The number of year expressions mentioned in relevant documents is an important feature which help us to differentiate between seasonal queries. For SOE queries, multiple year expressions with high frequency are expected. SSD queries in turn, are mostly characterized by one high frequency year expression. Finally, SHE queries have no year expression with high frequency. Based on this, we also consider: (3) **number of total year expressions**; (4) **number of distinct year expressions**; (5) **difference between the first and second frequent year expressions** (different between their frequency); and (6) **number of distinct year expressions** with frequency higher than 20.

4 EXPRIMENTS

4.1 Dataset and Experimental Setting

Our experiments were conducted on 300 Persian web queries (150 selected from Section 2, divided as follows: 41 SHE (Seasonal-related to historical events), 50 SOE (Seasonal-related to ongoing event) and 59 SSD (Seasonal-related to special days and traditions); plus 150 non-seasonal queries randomly selected from the query logs). Our dataset is publicly available[1]. To conduct our experiments, we make use of Hamshahri news dataset [15], which covers a wide range of news in Persian language, including politics, entertainment and sports, and resort to the set of queries introduced in Section 2. For each query, top-200 relevant documents were retrieved using Okapi BM25 retrieval model. As each document has a publication date this dataset suits our experiment. Time series were then generated

[1] http://dbrg.ut.ac.ir/SeasonalQueryDataset/

using documents publish date. To extract year expressions, we consider any number between 1990 and 2030 (Gregorian calendar), and 1300 to 1400 (Jalali calendar) a year expression. We used 10-fold stratified cross validation, and averaged the results over 10 runs. We used Random Forest for the classification and compared it with LibSVM, AdaBoost, and Naïve Bayes.

4.2 Feature Evaluation

In order to study the importance of our features we used information gain ratio (IGR) on training data. Auto correlation (0.371 of IGR) and seasonality (0.337) were the most important time series feature in terms of measuring the periodicity of time series. The distinct (0.337), seasonality (0.325) and also the total number of year expressions (0.319) are also important features to discriminate between the different seasonal categories. In contrast, some time series features like modality (0.124) randomness (0.048) were less discriminative.

4.3 Experimental Results

In order to classify queries into seasonal categories we use Random Forest (RF) classifier due to its properties like bagging and boosting. We compared its effectiveness against three baseline models: LibSVM, Naïve Bayes and AdaBoost. The results of our experiments are shown in Table 2. All the results are statistically significant when comparing RF classifier with each one of the baselines, with p-value < 0.05 using the matched paired one-sided t-test. A careful observation of the results led us to conclude that RF Classifier achieved the highest effectiveness with 0.887 F-measure which outperforms the 3 other classifiers.

Table 2: Performance of different classifiers

Model	Precision	Recall	F-measure
Random Forest	0.887	0.887	0.887
LibSVM	0.799	0.797	0.790
Naïve Bayes	0.820	0.757	0.757
AdaBoost	0.794	0.847	0.820

To better analyze the outcomes of our approach, we present in Table 3 the confusion matrix for the Random Forest classifier. As this table shows, instances of SSD queries (Seasonal-related to special days and traditions) were wrongly labeled as NS (non-seasonal queries). The main reason for that is the stable and non-periodic shape of time series for some of its queries. This can be observed in Figure 2 for the query "*Father's Day*", which despite being an SSD query, portraits a steady non-periodical shape.

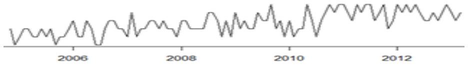

Figure 2: Time series built over top-200 relevant documents publish time retrieved for the query "Father's day".

Also, some queries from SOE (Seasonal-related to ongoing event) were wrongly categorized as SHE (Seasonal-related to historical events). Our exploration of the results, shows that, while SOE queries are characterized by multiple occurrences, top-200 retrieved documents are formed, mostly, by texts referring to a very specific episode of the event. For example, for the query "*Oscar*", most of the documents retrieved in the top-200 relate to 2012 when a Persian movie ("Separation") won the Oscar for the best movie. Yet this query is related to several all Oscar events. On the other hand, NS queries behaved well and 96% of its queries were correctly classified.

Table 3: Confusion matrix for the Random Forest classifier

Classified Real	NS	SOE	SHE	SSD
NS	144	2	2	2
SOE	1	40	5	4
SHE	2	2	34	3
SSD	7	1	3	48

5 CONCLUSIONS

Seasonal queries are a sub-type of temporal queries, characterized by a change of search intents over time. Understanding this seasonal behavior, may help search engines to provide better ranking approaches and to respond with temporally relevant results leading eventually into user's satisfaction. Ideally, search engines would have different retrieval strategies for any of the different categories, using this additional information to provide better responses for their users. In this paper, we proposed an approach for identifying different seasonal queries by using time series and content features. We show how users' behavior toward these queries are different. Random Forest classifier is used for classification and achieved 88.7% F-Measure. As part of future work, we plan to propose a ranking approach that use the proposed taxonomy to better rank the retrieved results. Although our approach is totally independent of any language, we plan to do the same study on an English dataset.

6 ACKNOWLEDGMENTS

This research was funded by Project "TEC4Growth - Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020" which is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

REFERENCES

[1] Campos, R., Dias, G., and Jorge, A. (2011). What is the Temporal Value of Web Snippets. In WWW-TWAW'11, pp. 9-16.

[2] Campos, R., Dias, G., Jorge, A., and Jatowt, A. (2014). Survey of Temporal Information Retrieval and Related Applications. In CSUR, 47(2). Article No.: 15.

[3] Ghoreishi, S., and Aixin, S. (2013). Predicting Event-Relatedness of Popular Queries. In CIKM'13, pp. 1193-1196.

[4] Goodwin, P. (2010). The Holt-Winters Approach to Exponential Smoothing: 50 Years Old and Going Strong. In The Int. Journal of Applied Forecasting, 19, pp. 30-33.

[5] Gupta D. and Berberich, K (2015). Temporal Query Classification at Different Granularities. In SPIRE'15, pp. 156-164.

[6] Hartigan, J. A., and Hartigan, P.M. (1985). The Dip Test of Unimodality. In The Annals of Statistics, 13(1), pp 70-84.

[7] Joho, H., Jatowt, A., Blanco, R., Naka, H., and Yamamoto, S. (2011). In NTCIR'11.

[8] Jones R. and Diaz, F (2007). Temporal Profiles of Queries. In TOIS, 25(3).

[9] Kanhabua, N., Nguyen, T., and Wolfgang, N. (2015). Learning to Detect Event-Related Queries for Web Search. In WWW'15, pp. 1139-1344.

[10] Kendall, M. G. (1948). Rank Correlation Methods.

[11] Metzler, D., Jones, R., Peng, F., and Zhang, R. (2009). Improving Search Relevance for Implicitly Temporal Queries. In SIGIR'09, pp. 700-701.

[12] Radinsky, K., Diaz, F., Dumais, S., Shokouhi, M., Dong, A., and Chang, Y. (2013). Temporal Web Dynamics and its Application to Information Retrieval". In WSDM'13

[13] Shokouhi M (2011). Detecting Seasonal Queries by Time-Series Analysis. In SIGIR'11, pp. 1171-1172.

[14] Zhang, R., Konda, Y.; Dong, A.; Kolari, P.; Chang, Y., and Zheng, Z. (2010). Learning Recurrent Event Queries for Web Search. In EMNLP'10, pp. 1129-1139.

[15] AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M. and Oroumchian, F., 2009. Hamshahri: A standard Persian text collection. Knowledge-Based Systems, 22(5)

[16] Kulkarni, A., Teevan, J., Svore, K. M., and Dumais, S.T. (2011). Understanding Temporal Query Dynamics. In WSDM'11, pp. 167-176.