Detecting Seasonal Queries Using Time Series and Content Features

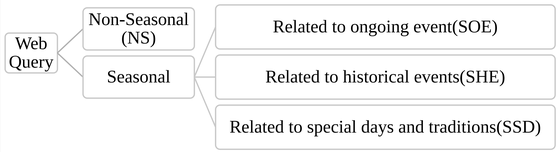
The 3rd ACM International Conference on the Theory of Information Retrieval (ICTIR 2017)
October 1-4 2017 | Amsterdam

Behrooz Mansouri	Mohammad Sadegh Zahedi	Maseud Rahgozar	Ricardo Campos
University of Tehran,Iran	University of Tehran,Iran	University of Tehran,Iran	Polytechnic Institute of Tomar, INESC TEC – LIAAD,Portugal
b.mansouri@ut.ac.ir	s.zahedi@ut.ac.ir	rahgozar@ut.ac.ir	ricardo.campos@ipt.pt

Introduction

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.

SEASONAL QUERIES Taxonomy



Category	Examples
NS	Online Books, Secure Passwords, Information Retrieval
SOE	Olympics, Oscar, US Presidential Election
SHE	September 11th, Iran Revolution, Bam Earthquake
SSD	Halloween, Father's Day, Christmas

Approach

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

Time Series Features

From time series built on top-200 relevant documents publish time

- Autocorrelation
- Seasonality
- Kurtosis
- Randomness Test
- Sum of squad errors
- Modality
- Mean

Content Features

From top-200 relevant documents

- Content Clarity
- Number of Total Year Expressions
- Number of Distinct Year Expression
- Difference Between the First and Second Frequent Year Expressions
- Number of Distinct Year Expressions with Frequency Higher than 20

Results

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

Performance of different classifiers

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

Confusion matrix for the Random Forest classifier

Literature cited

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

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.

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

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

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

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

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%

Dataset

- 300 Persian web queries (41 SHE, 50 SOE , 59 SSD and 150 NS)
- Manually classified by 4 professional editors (Fleiss Kappa statistics = 0.88)
- Hamshahri news dataset

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.

Download this poster

Further information

- http://dbrg.ut.ac.ir/
- www.ccc.ipt.pt/~ricardo/
- Or contact authors via email