ECAI 2016
G.A. Kaminka et al. (Eds.)
© 2016 The Authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms
of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/978-1-61499-672-9-1616

Collective Future Orientation and Stock Markets

Mohammed Hasanuzzaman[1] and **Wai Leung Sze**[2] and **Mahammad Parvez Salim**[3] and **Gaël Dias**[4]

Abstract. Web search query logs can be used to track and, in some cases, anticipate the dynamics of individual behavior which is the smallest building block of the economy. We study AOL query logs and introduce a collective future intent index to measure the degree to which Internet users seek more information about the future than the past and the present. We have asked the question whether there is link between the collective future intent index and financial market fluctuations on a weekly time scale, and found a clear indication that the weekly transaction volume of S&P 500 index is correlated with the collective intent of the public to look forward.

1 INTRODUCTION

Everyday, huge amounts of data are generated through the society's extensive interactions with technological systems, automatically documenting collective human behaviour on the Internet. The enormity and high variance of information that flow through the web opened new avenues for harnessing that data and associating online human activities with offline outcomes (real world social phenomena).

Analysis of web search queries, as logged by search engines such as Google, Bing, Yahoo and AOL has received increased attention in recent years. For example, health-seeking behavior in the form of online web search queries, which are submitted to Google search by millions of users around the world are used to improve early detection of seasonal influenza [2]. A strong correlation was found between the current level of economic activity in given industries and the search volume data of industry-based query terms [1]. However, it remains unexplored whether trends in financial markets can be anticipated by the collective temporal orientation of online users.

Temporal orientation refers to differences in the relative emphasis individuals place on the past, present and future, and is a predictive indicator of many human factors such as occupational and educational success, engagement in risky behavior, financial stability, depression and health [8].

In this study, we look into temporal orientation of search queries of online users. First, we developed a model on the basis of several linguistic features to automatically detect the temporal intent (oriented towards *past, present* and *future*) of search engine queries. We use this model to classify millions of web search queries as *past, present, future* or *atemporal* . Afterwards, we calculate the ratio of the volume of searches oriented towards the *future* to the volume of searches oriented towards the past, aggregate the ratios on a weekly time scale and call this value the *collective future intent index*. We find evidence that weekly trading volumes of stocks traded in S&P 500 are correlated with the *collective future intent index* of queries. Our results provide quantitative support for the recommendation that stock market predictions can be improved by the inclusion of the specific temporal dimension of online users.

2 EXPERIMENTS

2.1 Data Analysed

We use weekly aggregated transaction volumes of the S&P 500 and NASDAQ Composite stock index from 1st March, 2006 to 31st May, 2006 from Yahoo Finance[5]. AOL Query Logs distributed for non-commercial research purposes only by the AOL search engine [6] are used to compute the *collective future intent index*. This collection consists of approximately 20 million web search queries by 650,000 users over the three months from 1st March, 2006 to 31st May, 2006. The data set includes various information related to each query. We used only the query text issued by the user and the time at which the query was submitted for search.

2.2 Query Temporal Intent Classification

The idea behind query temporal intent classification is to determine whether there is a temporal dimension for users' information need [5]. This idea is pushed further in [3] and aims to determine whether the user is interested in information about the past, present or future when issuing a query or if his query has no temporal dimension. The task is to predict the temporal class $c \in \{$ *past, present, future, atemporal* $\}$ of a web search query with reference to its issuing date.

For the classification task, we use the gold standard data set composed of 400 queries (300 queries for training and 100 for test) released in the context of NTCIR-11 Temporalia shared task [4]. Examples of the form $< q, d, c >$ are as: *q*= who was martin luther, *d*= Jan 1, 2013, *c*= past; *q*= amazon deal of the day, *d*= Feb 28, 2013, *c*= present; *q*= stock market forecast tomorrow, *d*= Jan 1, 2013, *c*= future; *q*=number of neck muscles, *d*= Feb 28, 2013, *c*= atemporal.

We developed a methodology based on supervised learning technique exploring several features to label the temporal class of a query. The following features are used for the classification:

- **n-grams**: 1-2 token sequences. Features are encoded simply as binary indicators of whether the n-grams has appeared in the query.
- **timexes**: The mean difference between the resolved year date of time expressions and the issue year date. Time expressions are labeled and resolved via Standford's SUTime[6]. Specific features computed include the temporal difference itself, its absolute value and binary variables indicating any temporal expressions appearing in the query text that refer to the *future*, *past* or *present*.

[1] ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland

[2] Lamplight Analytics Limited, email: stephen.sze@lamplight.me

[3] Techno India Group, email: parvez.salim@gmail.com

[4] Normandie University, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France

[5] http://finance.yahoo.com/

[6] http://nlp.stanford.edu/software/sutime.html

- **Lexica**: The relative frequency of temporal categories in the query text based on a freely available temporal resource namely TempoWordNet[7]. The features are encoded as the frequency with which a word from a temporal category has (*past, present, future*) in the text divided by the total number of tokens in the text.
- **Time-gap**: If there is no mention of time expression inside q (timely implicit query), we consider the most confident year date obtained from freely available web service GTE[8]. Finally, the features recorded include the difference between the most confident returned year date and the issue year date, its absolute value, and confidence value returned by GTE. Some examples of extracted[9] year dates and confidence values are given in Table 1.

Query	Most confident Year	Confidence value
who was martin luther	1929	0.944
amazon deal of the day	2015	0.760
release date for ios7	2013	0.893
number of neck muscles	2014	0.708

Table 1: Examples of extracted year dates and confidence values

As for classifier, we used Support Vector Machines (SVM) from the Weka platform. We consider both a linear kernel (lSVM) and a radial basis function kernel (rSVM). From cross validation over the training data, we choose L1 penalization for lSVM and L2 ($\alpha||\beta||^2$) for rSVM. Our models are evaluated over the test data provided by the organizers. Overall results are presented in Table 2.

System	Atemporal	Past	Present	Future	All
lSVM	69.2	87.5	77.5	89.5	**80.0**
rSVM	68.5	84.7	74.5	86.4	78.0

Table 2: Classification accuracy for TQIC Task

Finally, the query temporal intent classification model is applied to time-tag AOL queries according to its underlying temporal intent. Afterwards, we calculate the ratio of total number of *future* oriented queries to the aggregated number of queries tagged as *past, present* for each week of our study period and consider these values as *collective future intent index* scores. For example, the *collective future intent index* for a given week (e.g. first week of April) is calculated as the ratio of the total number of searches with *future* intent to the total number of searches with *present* plus *past* intent for the same week (i.e. first week of April).

2.3 Time Series Correlation

After calculating the *collective future intent index* scores, we are concerned with the question whether it correlates with financial market fluctuations, in particular S&P 500 index traded volumes. To answer this question, we use the cross-correlation coefficient (CCF) in a similar fashion to [7]. The CCF estimates how variables are related at different time lags. The CCF value at time lag Δt between two time-series X and Y measures the correlation of the first time-series with respect to the second time-series in dependence of a time lag parameter Δt. The CCF is calculated as:

$$R_{xy}(t, \Delta t) = \frac{E[Y_{t+\Delta t}]E[X_t]}{\sigma[Y_{t+\Delta t}]\sigma[X_t]} \quad (1)$$

In this experiment, two time-series (X, Y) are change of aggregated traded volume of stocks and change of *collective future intent index* respectively on time with week granularity. The CCF values in dependence of different time lags are presented in Table 3.

	Time Lag (weeks)						
	-3	**-2**	**-1**	**0**	**+1**	**+2**	**+3**
S&P 500 index	-0.13	0.09	0.17	**0.41**	0.16	-0.12	-0.09
NASDAQ index	-0.17	-0.06	0.14	**0.34**	0.11	-0.16	-0.18

Table 3: Time lag-dependent cross correlation between weekly changes of transaction volumes of S&P 500 and NASDAQ Composite stock index and weekly collective future intent index changes.

It is evident from the results that the *collective future intent index* scores are correlated with the the aggregated volume of transactions for all stocks in S&P 500 constituents for a time lag of week (Δt = 0 week), i.e. the present week *collective future intent index* value is significantly correlated with present week transaction volumes of the S&P 500.

Moreover, we investigate the robustness of the *collective future intent index* by considering weekly transaction volume of NASDAQ Composite stock market index for the same period. Results in Table 3 illustrate the cross correlation between weekly aggregated volume of transactions and weekly *collective future intent index* changes. The same effects can be found for NASDAQ Composite, even if the CCF scores at time lag Δt = 0 week is smaller than for the S&P 500 constituents.

3 CONCLUSION

In this study, we have introduced the *collective future intent index* and its relation on the weekly transaction volumes of stock markets. One explanation for this relationship could be the *focus on the future supports economic success*. As future work, we are in the process of collecting long-term data, and will validate the results using long-term data. Afterwards, a Fuzzy Neural Network will be used to examine the predictive power of temporal orientation on the stock market.

REFERENCES

[1] Hyunyoung Choi and Hal Varian, 'Predicting the present with google trends', *Economic Record*, **88**(s1), 2–9, (2012).

[2] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant, 'Detecting influenza epidemics using search engine query data', *Nature*, **457**(7232), 1012–1014, (2009).

[3] Hideo Joho, Adam Jatowt, and Roi Blanco, 'Ntcir temporalia: a test collection for temporal information access research', in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion (WWW)*, pp. 845–850, (2014).

[4] Hideo Joho, Adam Jatowt, Roi Blanco, Hajime Naka, and Shuhei Yamamoto, 'Overview of ntcir-11 temporal information access (temporalia) task', in *NTCIR-11 Conference (NTCIR)*, pp. 429–437, (2014).

[5] Rosie Jones and Fernando Diaz, 'Temporal profiles of queries', *ACM Transactions on Information Systems*, **25**(3), (2007).

[6] Greg Pass, Abdur Chowdhury, and Cayley Torgeson, 'A picture of search.'.

[7] Eduardo J. Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes, 'Correlating financial time series with micro-blogging activity', in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, pp. 513–522, New York, NY, USA, (2012). ACM.

[8] Philip G Zimbardo and John N Boyd, 'Putting time in perspective: A valid, reliable individual-differences metric', in *Time Perspective Theory; Review, Research and Application*, 17–55, Springer, (2015).

[7] `https://tempowordnet.greyc.fr/`

[8] `http://wia.info.unicaen.fr/GTEAspNetFlatTempCluster_Server/`

[9] Extraction was processed 16th April, 2015 for illustration.