The Information Network: Exploiting Causal Dependencies in Online Information Seeking

Prasanta Bhattacharya
Department of Information Systems
National University of Singapore
prasanta@comp.nus.edu.sg

Rishabh Mehrotra
Department of Computer Science
University College London
r.mehrotra@cs.ucl.ac.uk

ABSTRACT

The Internet has emerged as a leading source of information about the world and its daily occurrences. Platforms like Wikipedia act as information conduits through which informational elements (e.g. topic pages) cater to the information seeking needs of users worldwide. While usage data from these informational elements help us to predict the information seeking behavior of users, especially in reaction to external news events, what has been largely ignored in past literature is the predictive value of the underlying informational network that connects these elements. In this study, we uncover causal linkages in information seeking behavior among related informational elements on Wikipedia. We demonstrate that incorporating this causal information leads to better predictions of page view counts of relevant Wikipedia pages, when compared to models that ignore such underlying causal linkages. We also provide additional evidence about the efficacy of our approach from the real world, by performing a judgment study with human annotators. This research is among the first to investigate and uncover the value of understanding the underlying relationships among informational elements.

Keywords

Granger Causality, Information Seeking, Recommendations

1. INTRODUCTION

The Internet has emerged as a leading source of information about the world we live in. Over 3 billion of the 7.2 billion people in this world, depend on the Internet to satisfy their informational needs in various forms [1]. Wikipedia, the largest free and multilingual encyclopedia on the Internet, has over 4 million articles on its English version as of July 2015 and experiences an addition of 1200 articles on average each day [2]. Other online platforms including Q&A sites (e.g. Stack Exchange and Quora), Blogs (e.g. Blogger and Tumblr) and Social Network Sites (SNS) (e.g. Facebook and Twitter) have also emerged as complementary sources of information for users on the Internet.

Given the rising popularity of information seeking behavior on the Internet, it has becomes crucial for platform owners to predict informational needs of its users in advance, and if possible, recommend sources to satisfy these needs. Predicting user interest in trends and events has been the subject of some recent analysis in this area [25]. However, predicting information seeking behavior and the subsequent recommendations are complicated by the fact that there often exists significant heterogeneity in user tastes online, and that the informational needs are quite dynamic in nature, showing fluctuations and seasonal trends.

The key intuition we exploit in the current study is that the Internet is composed of informational elements which cater to specific informational needs of its users. Examples of informational elements could be a Wikipedia page, a Tumblr blog or an entity on Freebase. By analyzing the usage patterns of these informational elements, we can not only make inferences about the informational needs of Internet users, but also make better predictions and recommendations that would benefit the users. While a number of recent studies have looked at Wikipedia [27], search queries [13] and information entities [20] to study related questions, none of them exploit the fact that these informational elements are often linked to each other and there often exist non-obvious causal interactions among these elements that can be exploited to make better predictions of information seeking behavior. For instance, an increase in information seeking propensity on the topic of "Ebola" might be a direct result of an increase in information seeking propensity of a related informational element e.g. "Recent Outbreaks". Further, this increase can potentially trigger a further increase in user interest and subsequent readership of related informational elements, like "World Health Organization (WHO)". Either way, what is important is to acknowledge that there exists an underlying network of relatedness between informational elements, that can be intelligently exploited to improve inferences about information seeking.

In the current study, we use longitudinal data on daily page views of Wikipedia pages for 4 popular world events to create a relatedness-graph of linked informational elements on Wikipedia. For each event, we uncover hidden and non-obvious causal links among related informational elements. We then show that by incorporating information about these uncovered links in our predictive model,

[1] http://www.internetworldstats.com/stats.htm
[2] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

CHIIR '16, March 13-17, 2016, Carrboro, NC, USA

DOI: http://dx.doi.org/10.1145/2854946.2854974

we are able to achieve lower error rates on predicting page view counts as compared to a baseline model which ignores such causal links. We also validate the recommendations made from our causal model using a real-world user judgment study which establishes the efficacy of our approach. Finally, we make a point about the temporal adaptiveness of our causal model by showing that the causal graphs we construct evolve in reaction to external events, by demonstrating distinct network characteristics for the different temporal phases (e.g. before and after an event).

2. RELATED WORK

There are two distinct lines of work related to the current research viz. extant work on information seeking and popularity prediction and research on drawing causal insights from observational data via statistical models. We next review relevant literature in both these areas.

Information Seeking & Popularity Dynamics:
Predicting trends in information access and content popularity finds application in many areas, including support facilitation, media advertising, content caching, revenue estimation, traffic management and macro-economic trends forecasting, to name a few. Some prior work [10, 30] show that there is a correlation in views that contents receive over time. There are, however, relatively fewer studies that forecast a value for the actual popularity of content. Lee et al. [16, 17] use survival analysis to evaluate the probability that a given content will receive more than some x number of hits. Jamali and Rangwala [15] predict the popularity of content using an entropy measure based on the "user-interest peak"and the "co-participatory network". Szabo and Huberman [30] present a linear regression model based on the number of views. This method was also applied by [32, 5] to create predictive popularity models in different feature spaces.

These existing information seeking models fail to capture the interdependencies between the different informational elements, something we focus on in the current piece of work.

Causality Models & Analysis:
Advancements in causal inference has led to the development of a plethora of new methods, both for causal structure learning and for making causal predictions (i.e., predicting the aftermath of interventions). Causal relations among time series data have been modeled with Granger causality [31], lagged correlation [19], Bayesian networks [36], among others. Granger causality measures a cause in terms of whether it passes Granger Test, i.e., whether a variable helps in predicating the future events for a related variable, beyond what can be predicted by using only historical events for the latter variable alone. Lagged correlation characterizes causal relations with the correlation between two time series shifted in time relative to one another. Causal Bayesian networks interpret causal relations with graphical models, in which the predecessors of a node are interpreted as directly causing the variable associated with that node.

A variety of causality mining techniques have been studied in past work. Chang *et al.* [11] propose a Granger causality based influence model for Twitter context summarization. Qui *et al.* [23] propose Granger graphical models as an effective and scalable approach for anomaly detection. Non-parametric generalization of the Granger graphical models called Generalized Lasso Granger (GLG) were proposed by Bahadori *et al.*[3] to uncover the temporal dependencies from irregular time series. More recently, Zong *et al.* [37] leveraged the causal and dependency structure among alerts sequences in data center monitoring systems. Finally, Granger causality has also been used to compute the cause and effect relationships for pairs of motion trajectories of a video [21].

In this work, we enrich the information seeking models with Granger causal dependencies and show that incorporating insights from predictive causal linkages between informational elements helps improve predictive performance and enables making better recommendations.

3. PROBLEM FORMULATION

Information seeking has emerged as one of the leading activities of users on the Internet today [7, 28]. The informational needs of a user can be very specific (e.g. searching for the capital of a particular country) or more navigational and exploratory (e.g. knowing more about the Ebola outbreak across the globe) [20, 6]. On the web, this information is supplied to users via a multitude of distribution platforms e.g. information repositories like Wikipedia, social media platforms like Twitter and Facebook, and Q&A sites like Quora and Stack Exchange. Each of these platforms, in turn, is an agglomerate of several informational elements which provide information on a multitude of topics and entities e.g. Pages on Wikipedia, Subjects on Stack Exchanges etc. More formally, we define Informational Elements as follows:

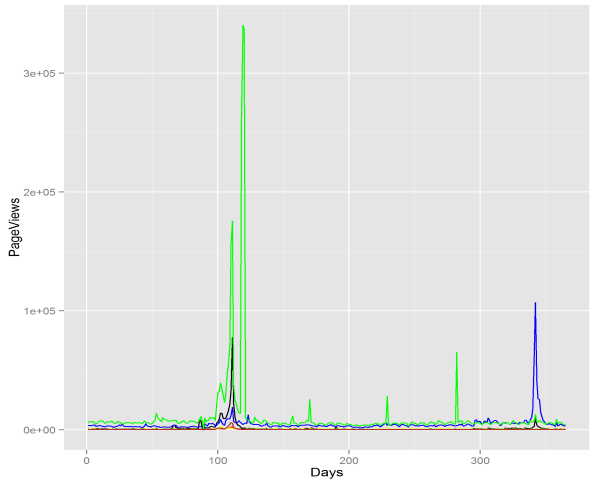
Definition: An *informational element* can broadly be defined as digital atomic units which have some informational value associated with them. Examples of informational elements could be various named entities, Facebook pages, Twitter handles, blogs, topics, Wikipedia pages, etc.

While the approaches discussed in this work are broadly applicable for any informational element, we make use of Wikipedia pages as the specific instance in this work. The usage activity logs from these informational elements provide a useful proxy to both measure and predict the informational needs of users.

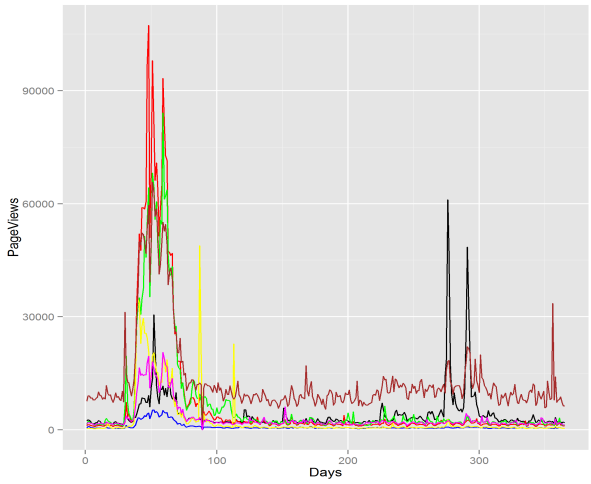
While there have been past work that predicts the popularity of topics and news events on social media platforms [33], we argue that the predictive power can be improved by exploiting causal linkages among related informational elements. The intuition behind this is that the popularity of a particular informational element, say after an important event, would increase popularity of related informational elements too. Hence, by incorporating popularity information from related elements, we should be able to make better predictions about the popularity of our focal informational element. Predicting popularity of informational elements is not only instrumental at illuminating our understanding of how individuals search for information on the Internet, but is also of key value to advertisers and platforms who wish to anticipate users' information needs and evolving preferences [24] following major events.

In the current study, we demonstrate our approach of identifying causal linkages, that are often non-obvious at times, using page-view logs from the English edition of Wikipedia [3]. We select a total of four Wikipedia pages corresponding

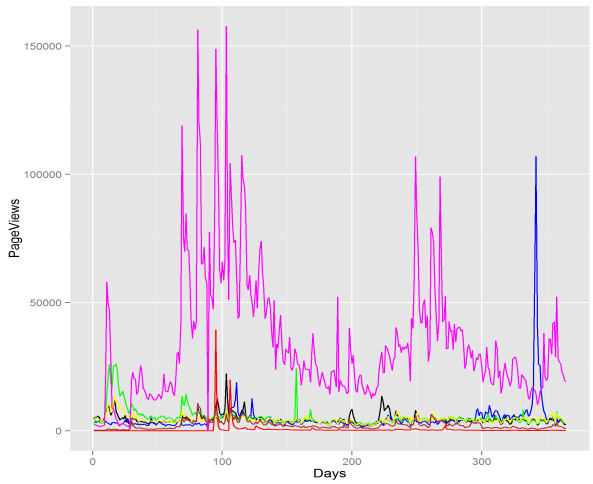
[3]https://en.wikipedia.org/wiki/Main_Page



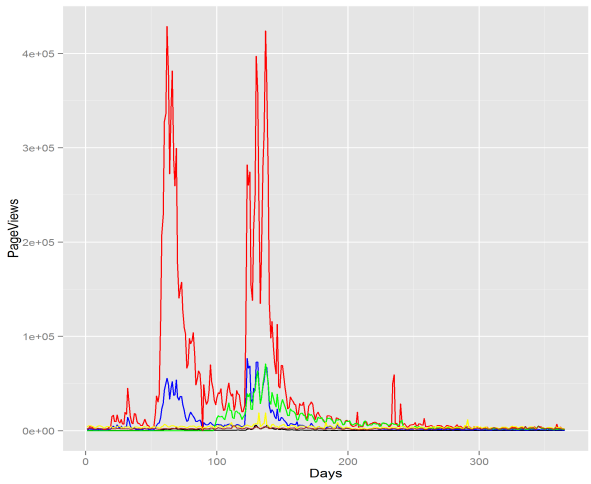
(a) Scotland Time Series



(b) Hamas Time Series



(c) ISIS Time Series



(d) Ebola Time Series

Figure 1: Time series of page views for the various informational elements considered in the different world events.

to major world events in recent history viz. The Scottish referendum on Independence, the Ebola outbreak, the rise and spread of ISIS militancy, and the rise of Hamas the Palestinian Islamic organization. Using an entity-tagging approach [9] that we describe in the following section, we identify a set of related Wikipedia pages for these four events. These are illustrated in Table 1 below. We then gather time series data on page view logs for each of these Wikipedia pages and visually inspect these to see if there is any evidence of the information seeking trends that co-evolve with each other.

Figures 1a, 1b, 1c and 1d illustrate the number of page views for each of the four events over a period of 1 year from June 1, 2014 to May 31, 2015. Interestingly, we find that for each of the focal events, while some related informational elements co-evolve in popularity, others don't. This poses an interesting problem of candidate selection when making predictions for the focal event as it is not obvious as to which informational elements might be useful to include in the predictive models. Specifically, we seek to answer the following two questions in this study, (i) "Can the predictive power of an informational element be improved by incorporating information from related informational elements?" and (ii), "Can we identify "causally" related informational elements from a set of all related elements?"

4. MODELING INTERACTION RELATIONSHIPS BETWEEN INFORMATIONAL ELEMENTS

In this section, we hypothesize that informational elements such as topic pages on Wikipedia are often causally related to each other in terms of information seeking patterns. We employ a Granger causality based approach to model these interactions between the elements.

4.1 Causality: Prior Art

Drawing causal conclusions for a set of observed variables

Event	Event Description	Informational Element (node ID)	Wikipedia Page
Scotland	The Scottish independence referendum on Scottish independence that took place in Scotland on Sept 18, 2014	Yes Scotland (1)	https://en.wikipedia.org/wiki/Yes_Scotland
		Edinburgh Agreement (2)	https://en.wikipedia.org/wiki/Edinburgh_Agreement_%282012%29
		Royal Bank of Scotland (3)	https://en.wikipedia.org/wiki/The_Royal_Bank_of_Scotland
		Scotland (4)	https://en.wikipedia.org/wiki/Scotland
		David Cameron (5)	https://en.wikipedia.org/wiki/David_Cameron
		Alex Salmond (6)	https://en.wikipedia.org/wiki/Alex_Salmond
ISIS	On 29th June 2014, ISIS proclaimed itself to be a worldwide caliphate.	Islamic State of Iraq (1)	https://en.wikipedia.org/wiki/Islamic_State_of_Iraq_and_the_Levant
		Syria (2)	https://en.wikipedia.org/wiki/Syria
		Iraq (3)	https://en.wikipedia.org/wiki/Iraq
		David Cawthrone Haines (4)	https://en.wikipedia.org/wiki/David_Cawthorne_Haines
		Islamic State (5)	https://en.wikipedia.org/wiki/Islamic_state
		Al-Qaeda (6)	https://en.wikipedia.org/wiki/Al-Qaeda
		Iraq War (7)	https://en.wikipedia.org/wiki/Iraq_War
		David Cameron (8)	https://en.wikipedia.org/wiki/David_Cameron
Ebola	The Ebola outbreak began in Guinea in December 2013 & then spread to Liberia & Sierra Leone	Ebola virus (1)	https://en.wikipedia.org/wiki/Ebola_virus
		Ebola Disease (2)	https://en.wikipedia.org/wiki/Ebola_virus_disease
		Ebola River (3)	https://en.wikipedia.org/wiki/Ebola_River
		WHO (4)	https://en.wikipedia.org/wiki/World_Health_Organization
		Ebola Epidemic in Africa (5)	https://en.wikipedia.org/wiki/Ebola_virus_epidemic_in_West_Africa
		Malaria (6)	https://en.wikipedia.org/wiki/Malaria
Hamas	Israeli air force attacks Hamas targets in central Gaza Strip	Israel (1)	https://en.wikipedia.org/wiki/Israel
		Gaza Strip (2)	https://en.wikipedia.org/wiki/Gaza_Strip
		Hamas (3)	https://en.wikipedia.org/wiki/Hamas
		State of Palestine (4)	https://en.wikipedia.org/wiki/State_of_Palestine
		Benjamin Netanyahu (5)	https://en.wikipedia.org/wiki/Benjamin_Netanyahu
		Fatah (6)	https://en.wikipedia.org/wiki/Fatah
		West Bank (7)	https://en.wikipedia.org/wiki/West_Bank
		Iron Dome (8)	https://en.wikipedia.org/wiki/Iron_Dome
		UN (9)	https://en.wikipedia.org/wiki/United_Nations

Table 1: World events and related informational elements on Wikipedia

from a given sample from their joint distribution is a fundamental problem. Statistical associations are often due to underlying causal structures [26]. Research in causal discovery has led to the identification of fundamental principles and methods for causal inference, including a complete algorithm, the PC algorithm, that identifies all possible orientations of causal dependencies from observed conditional independencies [29]. Identifying such causal relations helps uncover dependencies between variables which could be leveraged for different applications, in our case, making better predictions.

We first introduce the problem of causal inference on iid data, as with the case with no temporal structure. Let therefore X_i, $i \in V$, be a set of random variables and let G be a directed acyclic graph (DAG) on V describing the causal relationships between the variables. The *Causal Graphical Models* are usually thought of as joint probability distributions on the variables $X_1, ..., X_n$ with arrows indicating direct causal influences. The causal Markov assumption states that each vertex X_i is independent of its non-descendants in the graph, given its parents. Crucially, this links causal semantics, which is important for predicting how a system reacts to interventions, to something that has empirically measurable consequences. Given observations from a joint distribution, it allows us to test conditional independence statements and thus infer which causal models are consistent with an observed distribution, subject to a genericity assumption referred to as faithfulness.

We now turn to the case of time series data - which is of interest in the present study, and describe a popular framework to infer causal dependencies in temporal data.

4.2 G-Causality

Granger Causality [14] or "G-Causality", is one of the earliest methods developed to quantify the temporal-causal effect among multiple time series. It is based on two major principles: (i) the cause happens prior to the effect and (ii) the cause makes unique changes in the effect [4]. Such a formulation is based on the idea that a cause should be helpful in predicting the future effects, beyond what can be predicted solely based on their own past values. Specifically, a time series (or "page view count" series in the terminology of the present paper) x is said to *Granger cause* another time series y, if and only if regressing for y in terms of both past values of y and x is statistically significantly more accurate than doing so with past values of y alone.

More specifically, consider two vector autoregressive processes:

$$x_t = \Sigma_{i=1}^{\infty} a_{1i} x_{t-i} + u_{1t}; \quad var(u_{1t}) = \Sigma_1 \tag{1}$$

and

$$y_t = \Sigma_{i=1}^{\infty} b_{1i} y_{t-i} + v_{1t}; \quad var(v_{1t}) = \Gamma_1 \tag{2}$$

which can be viewed as linear projections of x_t and y_t on their own past values, which we denote as X_{t-1} and Y_{t-1}, respectively. The linear projection of x_t on both X_{t-1} and Y_{t-1} and of y_t on both X_{t-1} and Y_{t-1} can be obtained from the joint auto-regressive process:

$$x_t = \Sigma_{i=1}^{\infty} a_{2i} x_{t-i} + \Sigma_{i=1}^{\infty} c_{2i} y_{t-i} + u_{2t}; \quad var(u_{2t}) = \Sigma_2 \tag{3}$$

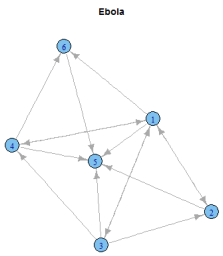
and

$$y_t = \Sigma_{i=1}^{\infty} b_{2i} y_{t-i} + \Sigma_{i=1}^{\infty} d_{2i} x_{t-i} + v_{2t}; \quad var(v_{2t}) = \Gamma_2 \tag{4}$$

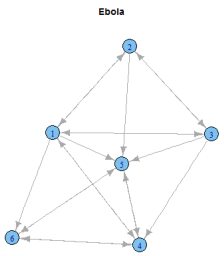
The variance Σ_1 represents the error in predicting the present value of x_t from its own past, while the variance Σ_2 represents the error in predicting the present value of x_t from the past values of both X_{t-1} and Y_{t-1}. If Σ_2 is less than Σ_1, then Y is said to cause X. This intuition is captured by the causal measure [14]:

$$F_{Y \to X} = ln(\frac{|\Sigma_1|}{|\Sigma 2|}) \tag{5}$$

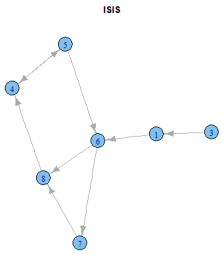
A similar measure of causality from X to Y can be computed by symmetry. However note that in general $F_{Y \to X} \neq F_{X \to Y}$, due to the directionality of the flow of time.



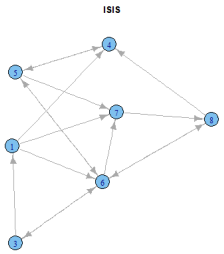
(a) Ebola event (threshold=0.01)



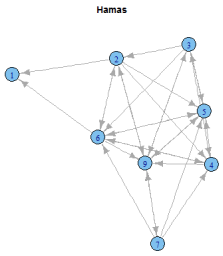
(b) Ebola event (threshold=0.05)



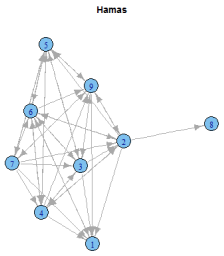
(c) ISIS event (threshold=0.01)



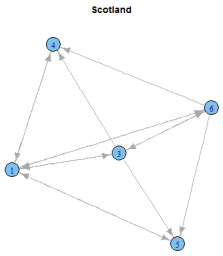
(d) ISIS event (threshold=0.05)



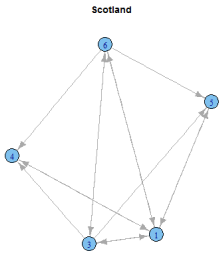
(e) Hamas event (threshold=0.01)



(f) Hamas event (threshold=0.05)



(g) Scotland event (threshold=0.01)



(h) Scotland event (threshold=0.05)

Figure 2: Causal network between various informational elements for the different events considered.

In spite of offering an enhanced degree of flexibility, parametric models of regression as formulated above suffer from potential exacerbation of performance inadequacy in cases where the true forms of correspondence between the response and the regressors may be non-linear. Non-parametric pairwise Granger causality is calculated as follows: given two point processes N_X and N_Y, a power spectral matrix S_{XY} is defined as the Fourier transform of covariance of two point processes N_X, N_Y , which is estimated using the multitaper function $h_k(t_j)$ [22]:

$$S_{XY}(f) = \frac{1}{2\pi KT}\Sigma_{k=1}^{K}\overline{N_X}(f,k)\overline{N_Y}(f,k)^* \quad (6)$$

where

$$\overline{N_X}(f,k) = \int_0^T h_k(t)exp(i2\pi ft)dN_X(t) \quad (7)$$

and S_{XY} is factorized by Wilson's algorithm [35] as follows:

$$S_{XY}(f) = H_{XY}(f)\Sigma_{XY}H_{XY}^*(f) \quad (8)$$

where H_{XY} is the transfer function which corresponds to the coefficients of an AR model, Σ corresponds to covariance matrix of error term of AR model and $*$ represents a conjugate transpose. Non-parametric pairwise Granger causality of $F_{N_Y \to N_X}$ for frequency f is finally calculated as:

$$F_{N_Y \to N_X}(f) = ln\frac{S_{XX}(f)}{S_{XX}(f) - (\Sigma_{YY} - \frac{\Sigma_{XY}^2}{\Sigma_{XX}})|H_{XY}(f)|^2} \quad (9)$$

We next describe our approach of using this notion of temporal causality in analyzing the interdependencies among the different informational elements.

5. EXPERIMENTAL SETUP

To investigate the implications of incorporating causal linkage information among informational elements, we consider a set of worldwide events as described in Table 1. For each of these events, we take a New York Times summary article which provides a descriptive summary of the event and use the content of the NYT article to extract the set of informational elements via entity extraction as described in Subsection 5.2 below. We next describe in detail the dataset construction and causal graph construction technique.

5.1 Dataset Description

The data used for this study is a time-series dataset on the number of daily page views for the English version of Wikipedia. Wikipedia not only allows its users instant access to information on virtually any topic of interest, but also provides usage metadata (e.g. page views, page edits etc.) through its periodic data dumps. The page view statistics we use for our study were collected from Wikimedia data servers [4] which were made more accessible through an external web application [5]. We collect daily page views for the period of June 1, 2014 to May 31, 2015 for the event pages on Wikipedia, as well as for the related pages as listed in Table 1 earlier.

[4]http://dumps.wikimedia.org/other/pagecounts-raw/
[5]http://stats.grok.se/

	1%		5%		10%		Common Baselines	
	G-Causal	**Shuffle**	**G-Causal**	**Shuffle**	**G-Causal**	**Shuffle**	**Shuffle (all edges)**	**No edges**
Scotland	**1693.8**	3536.1	**1693.8**	3550.4	**1693.8**	2484.4	4715.3	5001.9
ISIS	**1395.2**	1419.2	**1395.2**	1419.2	**1395.2**	1419.2	3405.2	3575.1
Ebola	**295.6**	302.02	295.6	294.8	**295.6**	302.0	6171.6	6152.5
Hamas	**3419.5**	3463.2	3419.56	3419.5	**3419.5**	3446.5	2652.1	2705.9

Table 2: Prediction estimates on the different events considered. RMSE values are reported for the proposed Causal approach and the different baselines considered. Row headers indicate the level of statistical significance considered.

5.2 Entity Extraction

The entity linking task aims at identifying all the small text fragments in a document referring to a particular entity contained in a given knowledge base, e.g., Wikipedia. The annotation is usually organized in three tasks. Given an input document, the first task involves discovering the fragments that could refer to an entity. Second, since an individual mention could refer to multiple entities, it is necessary to perform a disambiguation step, where the correct entity is selected among all possible candidates. Third and finally, discovered entities are ranked by some measure of relevance. More specifically, we use Dexter [8, 9] to link the events considered with entities. Dexter, in turn relies on DBPedia for entities and their type information.

The entities extracted via this technique are mapped to corresponding Wikipedia pages and the page view statistics are obtained for each of these entities. As discussed before, we treat these Wikipedia entities as the specific instance of informational elements and base our work on predictions and recommendations around these Wikipedia entities.

5.3 GCausal Graph Construction

We formalize our set of 4 focal events using a vector Event {Scotland, Ebola, ISIS, Hamas}. Each element $Event_i$ is associated with a list of related informational elements as listed earlier in Table 1. Drawing on our discussion of Granger causality in Sec. 4.2, we perform bivariate Granger casuality tests on every pair of informational elements in $Event_i$ and obtain a $n(Event_i) \times n(Event_i)$ causal adjacency matrix (CAM) for each focal element $Event_i$ where $n(Event_i)$ is the number of informational elements related to the focal event, including the informational element for the page itself. Each entry *(m,n)* in CAM represents the statistical significance of the non-parametric G-causality test between informational elements m and n. Next, we prune this adjacency matrix to remove all edges where the statistical significance is below certain level of significance. The resulting G-causal graphs for the four events at 1% and 5% levels of significance are shown in Fig. 2 where the node labels correspond to the IDs mentioned within parentheses alongside the informational elements in Table 1.

6. PREDICTION EXPERIMENT

In this section, we implement a G-Causality model based on the causal network identified, and fit it to the observed page-view data from Wikipedia. In order to emphasize the value proposition of incorporating the causal linkages, we baseline our results against that from a shuffle-test which replaces the causal predictors with a randomly selected, but informationally related predictor.

6.1 Predicting Page Popularity

In this section, we illustrate the predictive value of incorporating causal relationships between related informational elements as depicted in Fig. 2. For each informational element in $Event_i$, we choose the best predictor based on the G-causality test results as explained in Sec. 4.2. The best predictor for each target node is selected by comparing the F-statistics across the bivariate causality tests for every pair of nodes. We apply a time-series technique, namely, Vector Autoregressive Model (VARX) which captures dynamic feedback effects [12, 18, 1]. The model specification has been described earlier in Sec. 4.2.

This modeling approach allows us to explain the volume of page views for a particular informational element as a function of the volume of page views from past years of the same informational element, as well as the volume of page views of the "most related" informational element. The "most related" informational element is selected as the best predicting element, mentioned above. The root-mean-square of the estimation residuals are described in Table 2 above.

6.2 Baselines: Shuffle Test

We emphasize the benefits of uncovering causal predictors, using an edge shuffle test, similar to the one described in [2]. Specifically, we randomize the predictor nodes for each target node in our G-causal graph, by choosing randomly from a set of all candidate predictors including but not limited to the best predictor. We hypothesize that if there is no advantage to including best predictors from related informational elements, the shuffle test should not provide any significant reduction in predictive accuracy. The RMSE of model residuals from the shuffle test are provide alongside the RMSE of our G-causal model in Table 2 above. The common baselines include two cases viz. first, when the causal graph is not pruned based on the significance level and shuffling is performed on all edges, and second, when no causal graph is constructed and each variable is predicted only using an auto-regressive model.

6.3 Results & Discussions

As evident from our experiment results, incorporating information about related best predicting information elements provides an improvement in prediction accuracy over related but weak predicting information elements. Thus, while we contend that related information elements are useful predictors of popularity of the focal information element, it is important to identify the best predicting elements from the pool of all related elements. The choice of the best predicting element is often non-obvious and requires statistical causality tests for its identification. We show in Table 2 that the results from our causal prediction model outperforms results from the shuffle test model for all information elements, across almost all levels of significance.

	Causal 0.01	Causal 0.05	Shuffle	Content
	Relatedness			
Related	**70%***	63%	60%	67%
Somewhat Related	20%	23%	17%	30%
Not Related	10%	14%	23%	3%
	Interestingness			
Interesting	**73%***	67%	57%	63%
Somewhat Interesting	20%	23%	17%	20%
Not Interesting	7%	10%	26%	17%
	Informativeness			
Informative	**63%***	**63%***	50%	57%
Somewhat Informative	33%	27%	40%	37%
Not Informative	4%	10%	10%	6%

Table 3: Performance in terms of Relatedness, Interestingness & Informativeness for the Wikipedia page recommendation task. The results highlighted with * signify statistically significant difference between the proposed Causal recommendation framework and the best performing baseline using χ^2 test with $p \leq 0.05$.

7. RECOMMENDATION JUDGMENT STUDY

In addition to the prediction experiments, we evaluate the prowess of the proposed causal graph in making causal recommendations. We build the causal graph and use the methods proposed in Section 4, along with other baselines to generate Wiki page recommendations and perform a judgment study to evaluate the quality of recommendations. Next, we discuss the methodology and findings from the judgment study in detail.

7.1 Research Questions

To assess the quality of our casual graph for content recommendation, we performed crowd-sourced assessments with human annotators to seek answers to the following research questions:

RQ1: Relatedness: Are the recommended Wiki pages related to the original page? Relatedness is important since readers are unlikely to be interested in unrelated suggestions.

RQ2: Interestingness: Will readers be interested in exploring the recommended Wiki page given their original Wiki page visit? Interestingness is important since we are not trying to propose replacement or surrogates of the current page. Hence, a reader is likely to be interested in the suggestions if they are both related and novel.

RQ3: Informativeness: Are the recommendations intrinsically informative? Informativeness is an important characteristic since and it is preferable to avoid redundancy in recommendations, while at the same time suggest Wiki pages which provide some additional information to the reader who was interested in reading the original Wiki page.

7.2 Study Methodology

Given an event, we have a list of related informational elements and the Granger causal graph among these elements. The premise of this study is to show that these causal graphs can be used to improve content recommendations. In order to do so, we structure this judgment study in the following way: a user is shown contents from an initial wiki page (i.e. a base Wiki page) following which she is shown one recommended Wiki page based on the method being evaluated. We populate the set of recommend Wiki pages using the proposed Granger Causal graph as well as the baselines. The suggestions were labeled by judges who were recruited to participate in the judgment study. We used hidden quality control questions to filter out poor-quality judges. We had three judges in total, with each judge being shown a base Wiki page and asked to rate the recommended Wiki page on a number of measures evaluating the different aspects of recommendations. The procedure involving all the 30 informational elements, 4 different methods and the three measures yielded a total of 360 judgments.

The objective of this judgment study was to evaluate the quality of the recommendations, and answer RQ1 - RQ3 described above. As comparator methods, we generate and compare suggestions using the following techniques, which includes variations of the parameters in the proposed methods and some other methods used as baselines in the study.

- **Granger Causal graph based (2 variants):** Using the Granger Causal graph constructed, we take note of the directed relationships from the original Wiki page shown to the user and select the recommended page from the set of elements (Wiki pages) which had a directed link from the original wiki page in the causal graph constructed for the event (Fig. 2).

- **Content based:** Making use of content overlap between the base Wikipedia page and the recommended Wiki page, this baseline makes recommendations based on the most similar page to the current base Wiki page being viewed by the user. Intuitively, such recommendations should score well in terms of relatedness based measures and provide informative resources to users.

- **Shuffle Test based:** For this baseline, we make use of the linkage graph obtained via the shuffle test and recommend Wiki page which has a directed link from the original base Wiki page in the graph obtained in the Shuffle test as described in 6.2.

For every method, we show a maximum of 4 recommendations to judges and the judges were asked to judge the

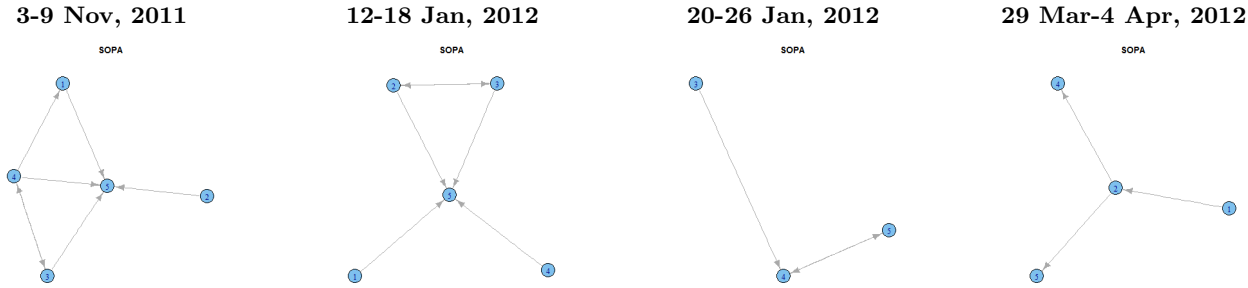


Figure 3: Analyzing the temporal evolution of causal linkages.

	1%				5%				10%			
	Pre 1	**Pre 2**	**Post 1**	**Post 2**	**Pre 1**	**Pre 2**	**Post 1**	**Post 2**	**Pre 1**	**Pre 2**	**Post 1**	**Post 2**
Degree	2.8	2.4	2	1.5	2.8	2.4	2	1.5	2.8	2.4	2	1.5
Closeness	0.06	0.064	.75	0.75	0.082	0.068	0.278	0.133	0.139	0.092	0.133	0.188
Betweenness	0	0	0	0	0.2	0	0.33	0.5	0.8	0.4	0.25	0.25

Table 4: Centrality measures for the different stages of SOPA event.

recommendations on the following dimensions on a three-point scale:

Relatedness: Suggestions are: (1) *Related*: all suggestions are related to the original Wiki page; (2) *Somewhat Related*: many suggestions are related to the original Wiki page; or (3) *Not Related*: most or all of the suggestions are not related to the original Wiki page.

Interestingness: Suggestions are: (1) *Interesting*: all suggestions are interesting given the original page; (2) *Somewhat Interesting*: many suggestions are interesting; or (3) *Not Interesting*: most or all suggestions are uninteresting.

Informativeness: Suggestions are: (1) *Informative*: all suggestions are informative given the original page; (2) *Somewhat Informative*: many suggestions are informative; or (3) *Not Informative*: most or all suggestions are uninformative given the contents of the original Wiki page.

Since most judges label largely disjoint sets of aspects, we do not report the standard Cohen's kappa for inter-annotator agreement. Instead, we report label agreement, which was 89.4%, 81.4% and 84.0% for relatedness, interestingness and informativeness respectively. This level of agreement demonstrates that judgment variance is quite small, and increases our confidence in the reliability of the judgments for evaluating our methods.

7.3 Findings

Table 3 shows the percentage of each response for the proposed methods and baselines in terms of relatedness, interestingness and informativeness.

Relatedness: The table shows that the causality based methods perform the best in terms of relatedness of the recommended Wiki page with the best performing method being Causal model with threshold 0.1. It is interesting to note that the content based recommendation performs better than Shuffle test based recommendation and Causal 0.5. This is not as surprising since a method that tries to find a related Wiki page which overlaps with the base Wiki page in terms of words, is bound to find a page which is very similar in terms of content. However, incorporating the causal aspect improves the score further to 70%.

Interestingness: Interestingness is a more important measure than relatedness given that one of the contributions of this user study is in devising techniques to better recommend Wikipedia pages to read. Like relatedness, we notice that the proposed causality based methods outperform all other baselines and with a bigger difference. Unlike Relatedness, both the causality based approaches beat the Content based and Shuffle test based recommendation baselines. This shows that incorporating causal linkages information while making recommendation indeed helps recommend more *interesting* content.

Informativeness: In terms of informativeness, the overall percentages obtained are less than those for relatedness and interestingness: 63% as compared to 70% & 73% but we do observe that the proposed causal based approaches outperform the baselines and hence recommend Wiki content which is more informative.

Overall, the findings of this analysis show that the Wiki page recommendations generated using our methods yield significant gains over the baselines in a number of important measures of recommendation value. This supports our claim that incorporating causal linkage information embedded in the information seeking behavior of the crowd helps uncover hidden insights which could be leveraged to make better recommendations of content.

8. TEMPORAL EVOLUTION OF CAUSAL LINKAGES

In the previous sections, we have established the importance of understanding causal linkages between related informational elements in predicting the information seeking

behavior of Internet users. However, we also hypothesize that these linkages are highly dynamic and often sensitive to major external events. To put it in different words, the causal networks for the focal events would evolve in a way that new linkages would emerge, while older linkages would lose strength and gradually disappear. To empirically investigate whether the causal graphs are indeed dynamic, we exploit a major World event to analyze whether the information seeking of event related pages show any short-run and long-run changes.

8.1 Event Context

On Jan 18, 2012, a number of major Internet-based organizations including Google, Wikipedia, Reddit etc. coordinated a series of protests against two proposed laws in the US Congress viz. the Stop Online Piracy Act (SOPA) and the Protect IP Act (PIPA). As part of the protest, some of the websites shut down their services and directed their users to a page displaying a protest message. This effect was clearly noticed, and within hours, other companies like Mozilla and Flickr joined in. The event triggered significant public participation with over 8 million people looking up their representatives on Wikipedia, and Twitter recording over 2.4 million anti-SOPA tweets. Giving in to popular opinion, both bills were finally removed from further voting by Jan 20, 2012.

8.2 Causal Linkages

Using our methodology as described in Sec.5.3, we constructed Granger causal graphs for the SOPA-blackout related informational elements, as shown in Fig. 3. In order to investigate the evolution of the causal graphs, we constructed these graphs at four different time stamps viz. (i) two months before the blackout (Pre 1) , (ii) a week before the blackout (Pre 2), (iii) a week after the blackout (Post 1), and (iv) two months after the blackout (Post 2). A simple visual inspection of the Figure 3 uncovers significant changes in the causal nature of linkages across the four periods. Beyond the change in network composition, we also perform centrality analysis [34] on the information elements in the graphs to verify if the overall centrality of the causal graph also changes over time. The centrality measures of degree, betweenness and closeness have been popularly used in the social networks literature as proxies to characterize the importance of the members of the social network. The results for the sociometrics are illustrated in Table 4 and confirm our hypothesis that both the composition of the causal graph, as well as the relative importance of the nodes change in response to major external events.

9. CONCLUSION & FUTURE WORK

Our research offers an early attempt at proposing a method to identify and incorporate causal linkages among informational elements on the Internet. The user access logs on information repositories like Wikipedia offer an invaluable source of data about the information seeking behavior of users. We demonstrate that such logs can be effectively exploited to uncover causal relationships among informational elements, that are not always obvious from a model-free analysis of the data. We highlight the incremental benefits to incorporating such causal information in our predictive model, over baseline approaches that ignore such causal linkages by randomizing the causal network. We then provide converging evidence from a judgment study where we asked human annotators to judge pairs of Wikipedia pages on Relatedness, Interestingness and Informativeness. Consistent with our predictions, we found that when users were exposed to pairs of pages that had causal linkages, they rated their experience more favorably as compared to those users who were exposed to a random pair of pages.

In addition to the predictive value of causal linkages, we also emphasize that this causal network among informational elements is not static, and is sensitive to major external events. Using the much popularized SOPA internet blackout as a test case, we demonstrated how the causal network changed at 4 different time periods before and after the event. An important recommendation we make based on this observation is that any predictive model for information seeking on the Internet must inherently be a dynamic one, and would need to be updated after major world events that are likely to influence the focal informational element.

10. REFERENCES

[1] G. Adomavicius, J. Bockstedt, and A. Gupta. Modeling supply-side dynamics of it components, products, and infrastructure: An empirical analysis using vector autoregression. *Information Systems Research*, 23(2):397–417, 2012.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15. ACM, 2008.

[3] M. T. Bahadori and Y. Liu. Granger causality analysis in irregular time series. In *SDM*, pages 660–671. SIAM, 2012.

[4] M. T. Bahadori and Y. Liu. An examination of practical granger causality inference. In *Proceedings of the SIAM International Conference on Data Mining, May*, pages 2–4, 2013.

[5] R. Bandari, S. Asur, and B. A. Huberman. The pulse of news in social media: Forecasting popularity. In *ICWSM*, pages 26–33, 2012.

[6] A. Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

[7] D. O. Case. *Looking for information: A survey of research on information seeking, needs and behavior*. Emerald Group Publishing, 2012.

[8] D. Ceccarelli, C. Lucchese, S. Orlando, R. Perego, and S. Trani. Dexter: an open source framework for entity linking. In *Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval*, pages 17–20. ACM, 2013.

[9] D. Ceccarelli, C. Lucchese, S. Orlando, R. Perego, and S. Trani. Learning relatedness measures for entity linking. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 139–148. ACM, 2013.

[10] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730. ACM, 2009.

[11] Y. Chang, X. Wang, Q. Mei, and Y. Liu. Towards twitter context summarization with user influence

models. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 527–536. ACM, 2013.

[12] M. G. Dekimpe and D. M. Hanssens. Sustained spending and persistent response: A new look at long-term marketing profitability. *Journal of Marketing Research*, pages 397–412, 1999.

[13] N. G. Golbandi, L. K. Katzir, Y. K. Koren, and R. L. Lempel. Expediting search trend detection via prediction of query counts. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 295–304. ACM, 2013.

[14] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[15] S. Jamali and H. Rangwala. Digging digg: Comment mining, popularity prediction, and social network analysis. In *Web Information Systems and Mining, 2009. WISM 2009. International Conference on*, pages 32–38. IEEE, 2009.

[16] J. G. Lee, S. Moon, and K. Salamatian. An approach to model and predict the popularity of online contents with explanatory factors. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 623–630. IEEE, 2010.

[17] J. G. Lee, S. Moon, and K. Salamatian. Modeling and predicting the popularity of online contents with cox proportional hazard regression model. *Neurocomputing*, 76(1):134–145, 2012.

[18] X. Luo. Quantifying the long-term impact of negative word of mouth on cash flows and stock prices. *Marketing Science*, 28(1):148–165, 2009.

[19] A. A. Mahimkar, Z. Ge, A. Shaikh, J. Wang, J. Yates, Y. Zhang, and Q. Zhao. Towards automated performance diagnosis in a large iptv network. In *ACM SIGCOMM Computer Communication Review*, volume 39, pages 231–242. ACM, 2009.

[20] I. Miliaraki, R. Blanco, and M. Lalmas. From selena gomez to marlon brando: Understanding explorative entity search. In *Proceedings of the 24th International Conference on World Wide Web*, pages 765–775. International World Wide Web Conferences Steering Committee, 2015.

[21] S. Narayan and K. R. Ramakrishnan. A cause and effect analysis of motion trajectories for modeling actions. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2633–2640. IEEE, 2014.

[22] A. G. Nedungadi, G. Rangarajan, N. Jain, and M. Ding. Analyzing multiple spike trains with nonparametric granger causality. *Journal of computational neuroscience*, 27(1):55–64, 2009.

[23] H. Qiu, Y. Liu, N. Subrahmanya, W. Li, et al. Granger causality for time-series anomaly detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1074–1079. IEEE, 2012.

[24] K. Radinsky, F. Diaz, S. Dumais, M. Shokouhi, A. Dong, and Y. Chang. Temporal web dynamics and its application to information retrieval. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 781–782. ACM, 2013.

[25] K. Radinsky, K. Svore, S. Dumais, J. Teevan, A. Bocharov, and E. Horvitz. Modeling and predicting behavioral dynamics on the web. In *Proceedings of the 21st international conference on World Wide Web*, pages 599–608. ACM, 2012.

[26] H. Reichenbach and M. Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1991.

[27] A. T. Scaria, R. M. Philip, R. West, and J. Leskovec. The last click: Why users give up information network navigation. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 213–222. ACM, 2014.

[28] C. Shah. Collaborative information seeking: understanding users, systems, and content. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 765–766. ACM, 2012.

[29] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.

[30] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[31] J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

[32] M. Tsagkias, W. Weerkamp, and M. De Rijke. News comments: Exploring, modeling, and online prediction. In *Advances in Information Retrieval*, pages 191–203. Springer, 2010.

[33] M. Tsytsarau, T. Palpanas, and M. Castellanos. Dynamics of news events and social media reaction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 901–910. ACM, 2014.

[34] S. Wasserman and K. Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[35] G. T. Wilson. The factorization of matricial spectral densities. *SIAM Journal on Applied Mathematics*, 23(4):420–426, 1972.

[36] C. Yuan, X. Liu, T.-C. Lu, and H. Lim. Most relevant explanation: properties, algorithms, and evaluations. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 631–638. AUAI Press, 2009.

[37] B. Zong, Y. Wu, J. Song, A. K. Singh, H. Cam, J. Han, and X. Yan. Towards scalable critical alert mining. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1057–1066. ACM, 2014.