

Retrieving Potential Causes from a Query Event

Suchana Datta*

University College Dublin
suchana.datta@ucdconnect.ie

Debasis Ganguly

IBM Research Europe, Dublin, Ireland
debasis.ganguly1@ie.ibm.com

Dwaipayan Roy

GESIS, Cologne
dwaipayan.roy@gesis.org

Francesca Bonin

IBM Research Europe, Dublin, Ireland
fbonin@ie.ibm.com

Charles Jochim

IBM Research Europe, Dublin, Ireland
charlesj@ie.ibm.com

Mandar Mitra

Indian Statistical Institute, Kolkata
mandar@isical.ac.in

ABSTRACT

Different to traditional IR, which retrieves a set of topically relevant documents given a user query, we investigate *causal retrieval*, which involves retrieving a set of documents that describe a set of potential *causes* leading to an *effect* specified in the query. We argue that the nature of *causal relevance* should be different to that of traditional *topical relevance*. This is because although the causally relevant documents would have partial term overlap with the ones that are topically relevant for a query, yet it is expected that a majority of these documents would use a different set of terms to describe a number of causes possibly leading to their effects. To address this, we propose a feedback model to estimate a distribution of terms which are relatively infrequent but associated with high weights in the *topically relevant* distribution, leading to potential causal relevance. Our experiments demonstrate that such a feedback model turns out to be substantially more effective than traditional IR models and a number of other causality heuristic baselines.

ACM Reference Format:

Suchana Datta, Debasis Ganguly, Dwaipayan Roy, Francesca Bonin, Charles Jochim, and Mandar Mitra. 2020. Retrieving Potential Causes from a Query Event. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, July 25–30, 2020, Virtual Event, China. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3397271.3401207>

1 INTRODUCTION

In this paper, we explore a novel direction of augmenting the classic ranked list of topically relevant search results with an additional list of *causally relevant* results. A causally relevant document is one that provides information on the likely causes leading to an event specified within the user query. As a concrete example, on submitted queries pertaining to a specific news event (e.g., ‘drop of pound’ or ‘housing crisis’), a causal search system retrieves potentially relevant information required to construct further analysis for the purpose of automated (or semi-automated with humans-in-loop)

*Research conducted during author’s affiliation at Indian Statistical Institute, Kolkata.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '20, July 25–30, 2020, Virtual Event, China

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-8016-4/20/07...\$15.00

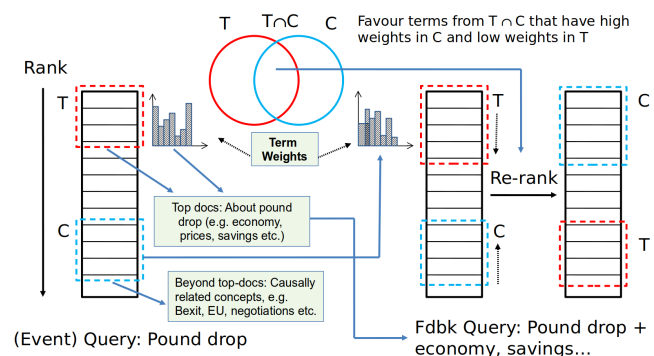
<https://doi.org/10.1145/3397271.3401207>

decision and policy making. Moreover, information extracted from causally related documents could also serve as the necessary *explanations* in order to support an automatically generated decision prescribing ways to eradicate a likely cause.

The study of causality in other domains, e.g. information extraction, is typically useful for finding explicitly mentioned causes, e.g. specified with typical patterns within a single sentence, such as ‘X leads to Y’ etc. However, the notion of causality that we address in the context of our work is more subtle; given a query regarding an event, the causal information need of the user is to find out the precursors of the event that lead to the event being taken place. Another point of difference of our work with existing ones is that we intend to retrieve causal information in the form of documents (query-document causal relations), the scale of which is typically much larger (and more subtle) than sentence level causality, such as ‘Heavy rain causes flood’.

Our work involves finding out the causally relevant information in an unsupervised way given a user-specified query describing an event with chronological dependency (e.g. ‘pound drop’ a potential cause of which is ‘Brexit’). In particular, we propose a pseudo-relevance feedback model based on the assumption that causally relevant documents are likely to have only a partial term overlap with the topically relevant ones, e.g., although the top-retrieved documents retrieved for the query ‘pound drop’ are likely not to contain terms, such as ‘Brexit’ or ‘EU’, it is likely that a number of documents beyond the top-ranks would indeed contain such terms. To address this, we propose a relevance feedback model that intends to estimate the distribution of causal terms that are relatively infrequent but associated with high weights in the topical relevance model. We observe that the association of these terms is significantly high in the topically relevant distribution, which

Figure 1: Schematics of FCRLM.



in effect leads to deviating away from topical relevance towards potential causal relevance.

The schematics in Figure 1 illustrates our idea that topical relevance (T) corresponds to the top-ranked documents, whereas the presence of causally relevant documents (C) may stretch further down the ranked list. In our proposed factored model, we use term distribution information from both these sets to build up a query with a broader aspect. Finally, we favour those documents which comprise terms that have higher weights in C and lower in T .

2 RELATED WORK

A number of existing approaches extract cause-effect patterns using lexical, syntactic, and more recently, semantic relations [3–5, 12], primarily taken from headlines or single sentences (refer to [1] for a survey). The cause-effect pattern approach was extended in [7], where a set of patterns are initially used to create a network of causes and effects, and then a relational embedding method (similar to TransE [2]) is used to jointly embed causes and effects. Causal IR works in the reverse direction, where a query describes a cause (e.g., current situation or proposed action) and the results provide a list of possible effects was studied in [9, 10].

3 FACTORED CAUSAL RELEVANCE MODEL

We estimate a causal relevance model θ_C as a function of the topical relevance model θ_T , which represents a set of terms that are directly related to the core topic of the query. This eventually expands the set of initial query terms because a small number of initial query terms is likely not to contain adequate information to help find the causally relevant terms (and eventually the documents). We estimate a standard *topical* relevance model, θ_T as:

$$P(w|\theta_T) = \sum_{i=1}^M P(w|D_i) \prod_{q \in Q} P(q|D_i), \quad (1)$$

where the weights $P(w|\theta_T)$ capture the co-occurrences between terms in the M top-retrieved documents, and query terms $q_i (\in Q)$.

We consider these estimated terms in θ_T has the ability of making the query more general, which according to our hypothesis is more useful to find causal relevance (as compared to the keyword query pointing to an effect event). Our intention is to balance the trade-off that terms in the causal model should, on one hand, be correlated well with the general concepts of θ_T , and should not be concentrated on solely the query (effect event) specific terms on the other. Thus, we formulate causal relevance model θ_C as:

$$P(w|\theta_C, \theta_T) = \sum_{i=1}^M P(w|D_i) \prod_{\forall t \ni P(t|\theta_T) \neq 0} \frac{P(t|D_i)}{P(t|\theta_T)}. \quad (2)$$

Specifically, Equation 2 reveals that in contrast to topical relevance model θ_T [11] estimation, we employ an *odds-ratio* between the probability of a term occurring within a document retrieved (with a more general query), and the probability of the same term estimated by θ_T (more specific effect event query). Inspecting the ratio term inside the product sign indicates that this way of estimation favours words that are: a) relatively infrequent in θ_T (i.e., terms not too specific to the effect event itself) through a decrease in the denominator; and b) frequent in the top-ranked documents retrieved with

Query	'Assassination of Osama-bin-Laden'
R_T	...buzz that Osama bin Laden may have died of typhoid in Pakistan last month...
R_C	The United States blames bin Laden and his Al Qaida network for the September 11, 2001, hijacked plan attacks on America ...

Table 1: Differences between two types of relevance.

the more general query, (i.e., terms related to more general aspects of the query, which as per our hypothesis constitutes a number of potential causes) through an increase in the numerator.

Finally, by substituting Equation 1 into 2, we get the final estimation model as:

$$P(w|\theta_C, \theta_T) = \sum_{i=1}^M P(w|D_i) \prod_{t \in \theta_T} \frac{P(t|D_i)}{\sum_{j=1}^M P(t|D_j) \prod_{q \in Q} P(q|D_j)}. \quad (3)$$

The qualifier 'factored' in the proposed model name indicates that the causal model is factorized into topical which then leads to estimating the causal one, as seen from Equation 3. Following the exposition of [8], the query model is linearly interpolated with the factored model in order to finalize the expansion term weights.

4 EVALUATION

Dataset Characteristics. Causality-based queries are associated with information needs with chronological dependency and having cause-effect relationship e.g. 'outbreak of a war between nations', 'major economic crisis' etc. Events for which there is a single cause, which is rather evident in nature (e.g., the cause is revealed in the article about the effect itself), are not interesting from the perspective of the causal retrieval task definition. In contrast to the direct cause-event relationships, we, in this work, are rather interested in those cases where pieces of causal relations are distributed across a number of different articles. The notion of *causal relevance* is determined by whether the information in a document relates to a potential cause of the effect specified in the given query.

Table 1 illustrates the differences between the two types of relevance for a sample query on the 'assassination of Osama bin Laden'. It can be observed that while the notion of *traditional relevance* (R_T) corresponds to the topic itself (the sample topically relevant document reports the possibility of Laden's death), the sample *causally relevant* (R_C) document comprises information about events that might have led to bin Laden's death (as shown in the bold-faced text).

Document Collection. Documents from the FIRE ad-hoc track English [6] was used as the target collection. It is comprised of 303,291 news articles crawled from *Telegraph India*¹.

Topics. As topics, we used our prior knowledge about the news events in the collection to select a set of events, such that for each event it can be reasoned that a number of factors could have been responsible in leading towards it. We exclude those cases where the causality factor is either too obvious (mentioned in the same document that describes the query event itself) or the number of such factors is too small in number (≤ 1). We also ensured that a query (topic)² is representative of an event that occurred during

¹<https://www.telegraphindia.com/>

²Available at <https://github.com/suchanadatta/Factored-Causal-RLM>

cause, lead, result, create, reason, generate, originate, spawn, effect, prompt, provoke, kindle, trigger, induce, inspire, promote, foster, beget

Table 2: Subset of causality-related words added to an initial query to explicitly seek causes of an event.

the period covered by the target collection, i.e. between 2001-2011. We compiled a total of 20 topics for our study.

Relevance Assessments. A *pool* of documents for manual relevance assessments in standard (i.e., topical relevance) IR is usually constructed by combining top-ranked documents retrieved by a number of systems (various IR models with different settings) [13]. In case of causal retrieval, this conventional way to construct a pool is likely not to work well because of two main reasons. Firstly, in contrast to topical IR, there exist no empirically well established model for causal IR (in fact, the purpose of developing manual assessments is to fill the void). However, relying on the proposed causal relevance model and a number of standard topical IR models (e.g. BM25, LM etc.) alone cannot ensure inclusion of the truly relevant documents in the pool. Secondly, unlike topical relevance judgement, assessors must have some prior knowledge on the event(mentioned in the query) without which it would be difficult to assess a causally connected event. To alleviate these issues, we treated causation finding for a topic as an exploratory task involving a series of query formulations and reformulations.

To aid our exploratory task, we used an interactive system that allowed *bookmarking* documents for future use (e.g., start exploring along a particular aspect of a potential cause of the main topic). At the end of the exploratory task, these bookmarked documents, being indicative of potentially relevant documents to establish the causal links with the query topic, were added to an assessment pool. Further, top-100 documents, retrieved with standard IR and feedback models, (specifically, LM, BM25 and RLM) were added to this pool. Documents from this pool were then assessed (binary judgments) using both prior knowledge on the topic and the knowledge gained during the exploratory session.

Methods investigated. Our first baseline is a standard IR model (specifically, linear smoothed language model) without feedback. Since our proposed approach is a factored feedback model, we employ the conventional relevance model based expansion technique RM3 [8] as our second baseline. As another baseline, we employ RM3 feedback to estimate θ_C with a modification: instead of assuming that the top-retrieved documents as relevant, we rather assume that the documents beyond the top-retrieved ones would be useful to estimate causal relevance. The intuition is as presented in Figure 1. For this, we swap the top M documents with a different set of M documents, $C_r = \{D_p, \dots, D_{p+M}\}$, where $p > M$ (i.e. an interval of documents of size M following the top- M). This baseline makes use of only the causal relevance assumption and it disregards information from the top-retrieved ones. We name this baseline ‘CRLM’ (RLM with causal relevance).

The next methodology that we investigate is to make queries more *specific* towards a causal information need. To illustrate with an example, the query ‘drop in pound value’ can be made more specific to a causal information need by *explicitly* adding cause-indicative keywords such as ‘causes’ or ‘reasons’ to the query. The purpose of reformulating the queries this way is to investigate

	MAP	NDCG	P@5	MRR
No-QE	0.3212	0.4909	0.2909	0.5432
RM3 ($M=10; T=40; \lambda=0.7$)	0.3441	0.5906	0.3300	0.6640
CRLM ($M=10; T=90; \lambda=0.5$)	0.2933	0.5280	0.3050	0.5944
No-QE-CSR	0.1687	0.2157	0.1747	0.2400
RM3-CSR ($M=20; T=40; \lambda=0.2$)	0.1758	0.2968	0.1894	0.3501
CRLM-CSR ($M=10; T=100; \lambda=0.1$)	0.1541	0.3280	0.2500	0.3783
FCRLM ($M=40; T=100; \lambda=0.3$)	0.3645	0.6197	0.4100	0.7219

Table 3: Comparison of retrieval effectiveness between FCRLM and a number of baselines. Improvements with FCRLM found out to be statistically significant with respect to all the baselines (t-test with $p < 0.05$).

if existing retrieval models could adequately address the causal information need if queries themselves explicitly indicate that information is sought on the causes related to an event and not the event itself, e.g. ‘reasons for the drop in pound value’ or ‘causes for the drop in pound value’.

For our experiments, we automatically constructed a set of such causality related keywords, which we add to make an initial query on an event more specific to seeking the causes for the event. A possible way of selecting these terms is to leverage synonyms of the word ‘cause’, as enumerated in Table 2. To find out if standard retrieval models on causally reformulated queries (with terms from Table 2) proves useful for retrieving causal information, we repeat all the previously mentioned baseline approaches on causally reformulated queries, which we denote by the suffix ‘CSR’ (abbreviation for Causality Specific Reformulation), e.g., ‘No-QE-CSR’ indicates LM based retrieval that uses queries augmented with the causality indicating terms shown in Table 2.

Parameter Settings. The common parameters in the methods investigated are a) the number of top-ranked documents, M , to consider as pseudo relevant, b) T , the number of top scored terms ($P(w|R)$ values), used for computing the KL divergence for reranking within a standard RLM setup [8], and c) the mixture parameter, λ , used in query likelihood. We optimize these parameters separately for each feedback method by grid search. The purpose of parameter tuning is not to claim that the best performing method also generalizes in the best possible way for other topics (typical of *learning* parameters in a supervised task), but rather to observe the best results achievable by each individual method. Grid search also ensures that the comparisons between the unsupervised methods are fair. Optimal results for each method, along with the optimal parameter values, are shown in Table 3.

Results. Table 3 shows comparison between the results obtained with the proposed method and the baselines. We enlist a number of observations as follows. Firstly, it can be noted that a standard (topical) feedback approach, such as RM3, results in a marginal improvement over the initial retrieval step (No-QE). This shows that applying off-the-shelf relevance feedback approaches may not prove effective in retrieving causally relevant information. So applying term expansions to diversify the query in order to find causally relevant terms may not help either because, as per our hypothesis, such terms are likely to have low likelihood of being generated

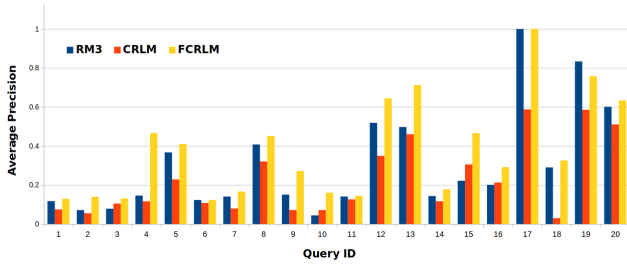


Figure 2: Per-query performance analysis in terms of average precision. Difference between blue and yellow bars indicates improvement of FCRLM over RM3 and CRLM.

from top-ranked documents only (which is the assumption in a standard feedback model).

Secondly, it can be seen that CRLM with its relatively simple heuristics of using pseudo-relevant information from documents deeper down the ranked list, is not able to improve results, which means terms not too related to the main information need (terms from mid-ranked documents) are rather more likely to be topically non-relevant than being causally relevant.

Thirdly, the approach of making a query more specific by explicitly adding the cause indicating terms (Table 2) usually contributes to decreasing retrieval effectiveness for all the above mentioned baseline approaches. This can be attributed to the fact that most causally relevant documents are likely not to explicitly mention the potential causes for an event, through the use of cause indicating words, e.g. it can be noted that the causally relevant documents in the example on Bin Laden’s assassination do not contain words such as ‘cause’, ‘reason’ etc. constituting Table 2.

From Table 3, we can further observe that the proposed FCRLM method turns out to be significantly better than all baseline techniques in terms of various evaluation metrics. From the improvement in precision at rank 5 and reciprocal rank based metric, it can be concluded that FCRLM is competent in retrieving more relevant documents at top ranks. Further, the improvement in average precision and cumulative gain based measurements indicate the overall proficiency of FCRLM in terms of all retrieved documents. These observations confirm the hypothesis that terms that are relatively infrequent in the topical feedback model, but at the same time, having a relatively high likelihood of occurrence within top-retrieved documents during a second step feedback, estimate the best notion of causal relevance. The factored nature of FCRLM and the use of the odds-ratio between the two factors prove to be useful to retrieve documents that are not related to the query directly, but rather represent the potential set of causally related precursors.

In addition to the aggregate results, we also present the per-query results in Figure 2, which shows that the improvements obtained with FCRLM are mostly consistent across a number of topics.

Analysis. We now present a qualitative comparison between the retrieval effectiveness of RM3 (best performing baseline) and FCRLM by enlisting the top set (by highest weights) of expansion terms selected by each techniques in Table 4 for a topic from our dataset. The topic alludes to a semi-political scandal that resulted in the resignation of Shashi Tharoor, a member of Indian parliament. *Lalit Modi*, the then *IPL* chairman, *tweeted* that Shashi’s friend *Sunanda*

Query Event	Shashi Tharoor resigned as member of parliament
RM3	shashi, tharoor, resign, franchis, controversi, kochi, congress, consortium, junior, kerala, modi, lalit, sunanda, minist, pushkar, rendevz, foreign, kolkata, sweat, ipl
CRLM	tharoor, shashi, wife, resign, junior, lalit, speech, minist, foreign, bid, statement, associ, question, embarrass, educate, cricket, kochi, prime, alleg, ask
FCRLM	tharoor, shashi, resign, pushkar, sunanda, ipl, kochi, tweet, consortium, lalit, equiti , kerala, modi, controversi, bjp , twitter, affair , kolkata, foreign, corrupt

Table 4: Top 20 expansion terms selected by RM3, CRLM and FCRLM for a sample query. Causally relevant terms exclusively estimated by FCRLM only are bold-faced.

Pushkar (with whom Shashi was allegedly having an *affair*) received *free equity* by team *Kochi*. It is seen from the highlighted words of Table 4 that FCRLM was successful in estimating terms that are relevant to the chain of events, such as ‘equiti’, ‘affair’, ‘corrupt’ etc., which its counterpart, RM3, was unable to find.

5 CONCLUSIONS AND FUTURE WORK

We proposed a different notion of ad-hoc search, where the information need is not directly related to the core topic of a query, but rather requires to find information that could have led to the query event. We constructed a collection for causality based IR research and proposed an unsupervised pseudo-relevance feedback algorithm that relies on a simple yet effective heuristic that causally relevant terms are not directly related to the core topic of the query. To leverage such terms, we rely on high term sampling probabilities from documents that are deeper down a retrieved list of topically relevant documents, and high co-occurrence likelihoods with the query terms to filter out potential noise.

As future work, we intend to explore ways of constructing causal chains of events in a recursive manner, i.e., instead of outputting a ranked list of documents we intend to extract events from the retrieved articles, treat them as queries in turn, and retrieve a list of further causes.

REFERENCES

- [1] Nabiha Asghar. 2016. Automatic Extraction of Causal Relations from Natural Language Texts: A Comprehensive Survey. *CoRR* abs/1605.07895 (2016).
- [2] Antoine Bordes et al. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Proc. of NIPS’13*. 2787–2795.
- [3] Du-Seong Chang et al. 2005. Causal Relation Extraction Using Cue Phrase and Lexical Pair Probabilities. In *Proc. of First IJCNLP*. 61–70.
- [4] Eduardo Blanco et al. 2008. Causal Relation Extraction. In *LREC 2008*.
- [5] Hashimoto et al. 2014. Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features. In *Proc. of ACL’14*. 987–997.
- [6] S. Palchowdhury et al. 2011. Overview of FIRE 2011. In *Proc. of FIRE’11*. 1–12.
- [7] Se. Zhao et al. 2017. Constructing and Embedding Abstract Event Causality Networks from Text Snippets. In *Proc. of WSDM’17*. 335–344.
- [8] N. A. Jaleel, J. Allan, W. B. Croft, F. Diaz, L. S. Larkey, X. Li, M. D. Smucker, and C. Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Proc. TREC ’04*.
- [9] Emre Kiciman. 2018. Causal Inference over Longitudinal Data to Support Expectation Exploration. In *Proc. of SIGIR ’18*. 1345–1345.
- [10] Emre Kiciman and Jorgen Thelin. 2018. Answering What If, Should I, and Other Expectation Exploration Queries Using Causal Inference over Longitudinal Data. In *Proc. of DESIRES’18*. 9–15.
- [11] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *Proc. of SIGIR’01*. 120–127.
- [12] Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning Causality for News Events Prediction. In *Proc. of WWW’12*. 909–918.
- [13] Ellen Voorhees and Donna Harman. 2000. Overview of TREC-8. In *Proc. of TREC-8*. 1–23.