Detecting significant events in news corpora

ABSTRACT

Forthcoming.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]

General Terms

Keywords

Event detection

1. INTRODUCTION

In this paper, we present a method for discovering significant events in news corpora. We propose a novel unsupervised method for event detection and introduce a test collection and evaluation methodology that can be used for future research.

2. USE CASES

Imagine a Wikipedia editor creating the 'year' page for a particular year (e.g., https://en.wikipedia.org/wiki/1988). These pages contain lists of important events that occurred during that year. Given a corpus of news articles (or other timestamped collection), one use case is to show the editor a list of major events that happened during a given year (or other time period) to facilitate the creation or validation of these pages.

A more general use case might be to summarize events in news corpora independent of Wikipedia. In this case, the user is looking for a list of major events that occurred in a given time span.

While newswire feeds are obviously amenable to this type of analysis, other timestamped document collections or streams (e.g., Twitter, Blogs) might also be useful.

3. WHAT IS AN EVENT?

The concepts of 'event detection' or 'event extraction' are common in the IR, NLP and ML literature. A variety of different definitions and operationalizations of 'events' are commonly used. In this paper, we are concerned with the retrospective identification of large-scale events in historical news corpora. A number of studies have used the concepts of "significant," "seminal," or "newsworthy" events to describe such large-scale events.

4. RELATED WORK

TBD

5. OUR MODEL

Retrospective event detection methods are generally of two types: document-centric or feature-centric. In document-centric event detection, documents are first clustered based on their content, then interesting temporal signals are identified from the clusters. In feature-centric event detection, the temporal distribution of individual terms are used as signals to identify events. Terms are then clustered into event models.

This study proposes a novel approach to feature-centric retrospective event detection. The model works as follows:

1. We construct a temporal index from the document collection. The index consists of document frequencies for each term in the vocabulary at a particular interval (e.g., hour, day, week).

2. For each term in the vocabulary, we construct a time series based on the distribution of the term frequencies over time. We then calculate the first-order autocorrelation (ρ) of the time series. Terms with high temporal dependency will have very high (1) or very low (-1) ρ. Terms with ρ values closer to 0 have low temporal dependency.

3. We assume that most terms in the vocabulary are not indicative of events. We assume that $\rho \sim N(\mu, \sigma)$ and that event-indicating terms are generated by a different process than non-event indicating terms. We test the hypothesis that term t is drawn from the non-event indicating distribution. If $p-value < \alpha$, reject the null hypothesis.

4. For all terms drawn from the event-indicating distribution, calculate the temporal mutual information $I_t(x; y)$ between terms.

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. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR '2016 Pisa, Italy
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

5. Construct a matrix of summary temporal mutual information between terms (e.g, maximum, average, minimum, variance). Use non-negative matrix factorization (NMF) to find the top k factors

6. Use the NMF factor weights to create the event model.

The resulting event model can be evaluated qualitatively (ala topic models) or used as a synthetic query to retrieve information about the event in the document collection or in an external collection, such as Wikipedia.

6. EVALUATION

There are three primary concerns in the evaluation of this approach:

1. Did the model find all of the significant events in the corpus? To assess this, we need a ground truth containing a comprehensive list of events in the corpus.

2. How many of the identified 'events' are truly events? In this case, we need to evaluate whether the events identified by the system correspond to events in the corpus.

3. How accurate are the identified date constraints? To address this question, we need the start and end dates associated with each identified event.

For this study, we develop an evaluation methodology based on Wikipedia 'year' pages supplemented with the manual assessment of candidate events returned by the systems.

6.1 Document collection

News articles from the years 2004 and 2005 in the New York Times Annotated Corpus are used for evaluation with stop words removed.

6.2 Wikipedia year pages

Using the Wikipedia 'year' pages, we construct an initial ground-truth of events. For each listed event, we extract the description and associated dates. Each identified event is then manually reviewed by two assessors to 1) assign an importance level to the event (described below) and 2) to identify the Wikipedia page that best represents the event.

A central concept in the evaluation of the system is event *importance*. One goal of the ground truth construction is to identify major events from the perspective of US news media. We define the following three categories:

1. Major: News articles about this event are likely to be found on the front-page. Interest in the event spans more than one week.

2. Moderate: News articles about this event may be found on the front-page. Interest in the event spans less than one week.

3. Minor: News articles about this event are unlikely to be found on the front-page. Interest in the event may only span a day or two.

Each assessor is asked to review the list of events, assign an importance level to the event, and identify the best Wikipedia page that describes the event, if present.

One obvious limitation of this approach is the assumption that Wikipedia contains descriptions of all events or that the year pages are comprehensive. We address this limitation through the qualitative analysis of events returned by the event detection systems, described next.

6.3 Manual assessment

For evaluation, each system returns the top N candidate events. For each event, the system returns an estimated event date and the top K Wikipedia pages that best represent each event. Two assessors are tasked with manually determining whether the candidate events represent *true* events and to collect additional details about the events for final evaluation.

For each event returned by the systems, two assessors are be tasked with reviewing the event models, results, and dates to assess the following:

1. Is this a "real" event (e.g., do you think the model is describing an event that really happened)?

2. If yes, what are the estimated constraints of the event (start/end dates)

3. Is the event represented in Wikipedia?

4. If yes, what is the best URL for the event?

5. If yes, how relevant are each of the results by the system (0=not relevant, 1=relevant, 2=the page is dedicated to the event)

6. What is the event importance (major/moderate/minor)?

6.4 Metrics

For each system, we can calculate recall and precision based on the combined lists of events identified by the Wikipedia editors and events identified by the systems. How many of the true 'major' events did the system identify? How many of the identified events are true 'major events'? Using the top K Wikipedia pages, we can also calculate IR metrics such as NDCG (how highly ranked are the best URLs for a particular event with respect to the event model)?

7. CONCLUSIONS

TBD

8. ACKNOWLEDGMENTS

This section is optional

APPENDIX