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. 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.

3. RELATED WORK

Forthcoming.

4. 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.

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.

5. EVALUATION

There are three primary concerns in the evaluation of this approach to event detection. First, did the model find all of the significant events in the corpus? To assess this, we would need a ground truth containing a comprehensive list of events in the corpus. Second, how many of the identified 'events' are truly events? In this case, we would need to evaluate the results returned by the system. Third, how accurate are the identified date constraints? To address this question, we would need the start and end dates associated with each identified event.

SIGIR '2016 Pisa, Italy

5.1 Manual review of events in Wikipedia

For this study, we develop an evaluation methodology based on Wikipedia. Wikipedia contains retrospective information about events themselves, as well as 'year' pages (e.g., `https://en.wikipedia.org/wiki/1988`) that include summary lists of important events. From the Wikipedia 'year' page we construct a ground-truth of events for the year. For each listed event, we extract the description, start date, and end date. Each Wikipedia 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. Entries that indicate events that were not newsworthy at the time (e.g., "Al-Qaeda is formed by Osama bin Laden") are not considered. While some events have dedicated pages, others are described in pages associated with individual entities and some have no representation in Wikipedia at all. A central concept in the creation of the ground truth is event "importance." The model proposed in this study is concerned primarily with significant or seminal events. One goal of the manual classification is to identify only major events from the perspective of US news media and the public at the time the event occurred. 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 briefly 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.

The resulting data serves as ground truth for the first part of our evaluation. The event detection system returns the top 50 event models and associated Wikipedia URLs. We can then evaluate system recall based on the events identified in the Wikipedia year page. Of course, Wikipedia does not provide complete coverage of all possible events. It is likely that the event-detection systems will identify valid events not included in Wikipedia. This is addressed in the next section.

5.2 Assessment of events identified by the system

The event detection systems return a set of event models for the top 50 events, represented as sets of terms and their weights. For each identified event, the systems also return the top 10 Wikipedia pages and a prediction of the event constraints as a set of start and end dates.

Each of these event models and associated Wikipedia pages are manually evaluated. For each event, two assessors will 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, how relevant are each of the 10 returned results (0=not relevant, 1=relevant, 2=the page is dedicated to the event)

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

The first test collection can be used to approximate system recall. The second test collection can be used to calculate a ranked metric, such as NDCG, for the returned Wikipedia pages. Another metric needs to be developed for the event constraints.

6. CONCLUSIONS

7. ACKNOWLEDGMENTS

This section is optional

APPENDIX