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 need a ground truth containing a comprehensive list of corpus events. Second, how many of the identified 'events' are truly events? In this case, we need to evaluate the results returned by the model. Third, how accurate are the identified date constraints? To address this question, we need the start and end dates associated with each identified event.

SIGIR '2016 Pisa, Italy

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, and end dates. Each Wikipedia event is then manually reviewed by two assessors to identify 1) the event category, 2) importance within the category (major, moderate, minor), and 3) to identify the Wikipedia page most representative of the event. If the event entry indicates something that was not newsworthy at the time (e.g., "Al-QaedaÊis formed byÊOsama bin Laden"), it is removed. While some events have dedicated pages, others are described in pages associated with individual entities and some have no representation in Wikipedia at all. The resulting data serves as ground truth for the first part of our evaluation.

The event detection systems return a set of event models for the top 50 events in the primary corpus. These event models are then used as queries against Wikipedia, returning the top 10 Wikipedia pages that match each event model. The output of the event detection system is compared to the list of events from the ground-truth. We calculate recall, precision, and NDCG.

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. To address this, the event models and returned Wikipedia pages are manually evaluated. Assessors consider the event model and constraint dates to judge each event (0 not an event, 1 is an event). For each true event, the returned Wikipedia pages are judged on a 3 point scale (0 not relevant, 1 relevant, 2 highly relevant).

6. CONCLUSIONS

7. ACKNOWLEDGMENTS

This section is optional

APPENDIX