TOPIC DETECTION AND TRACKING WITHIN SOCIAL NETWORKS

Supervisor
Dr. Ahmed Rafea

Nada Ayman
OUTLINE

- Introduction
- Methods for Topic Tracking
- Methods for Topic Detection
- Evaluation Metrics
- Challenges
- Goal
- References
INTRODUCTION

- Over the past few years the Internet has not only become the most important source of information, but also a key-player in event formation.
- The open community of publishing news and information made it an important indication for the pulse of the society.
INTRODUCTION

- Social networks have become a very important source of information and recently a source of creating information.
- Blogs, Twitter and Facebook, have played a great role in near past and current events all over the world.
- For all this, it was very important to have a system that can extract these information without human intervention.
WHAT’S TDT?

- The basic idea originated in 1996, when the Defense Advanced Research Projects Agency (DARPA) realized it needed technology to determine the topical structure of news streams without human intervention.

- A new area of research in information retrieval (IR) has developed over the past four years called Topic Detection and Tracking (TDT).
What’s TDT?

- Topic detection involves detecting the occurrence of a new event such as a plane crash, a murder, a jury trial result, or a political scandal in a stream of news stories from multiple sources.

- Topic tracking is the process of monitoring a stream of news stories to find those that track (or discuss) the same event as one specified by a user.
Tasks of TDT

- **Five tasks of TDT:**
  - **Story segmentation**
    No need for it if we have already the data as documents.
  - **Topic detection**
    Builds a set of clusters, each contains stories about the same topic. It assigns a story to one or more possible cluster.
  - **Topic tracking**
    is the selection of a certain cluster specified by one or more example stories
  - **First story detection**
    detects a story discusses unknown topic. It generates a new cluster for this topic.
  - **Link detection**
    a kernel function which established if two stories are linked or not.
TASKS OF TDT (Contd)
DATA PREPROCESSING

- Identify individual words and reduce the typographical variation. (Tokenization)
- Remove non-informative words. (Stop-words removal)
- Reduce morphological variation. (Stemming)
- Compute the term-weights. (Using TFIDF or other models)
- Build the vector.
METHODS OF TOPIC TRACKING

- Vector Space Model
- K-Nearest Neighbor (KNN)
- Language Models
- Use of semantical and contextual information
- Semantic classes
VECTOR SPACE MODEL

- The corpus is represented as $D = \{D_1, D_2, \ldots, D_m\}$ with $m$ documents and a term space $T = \{t_1, t_2, \ldots, t_n\}$ with $n$ terms.
- Each row contains relative significance of a term in a document which commonly represented by TFIDF weight.
Vector Space Model (Contd)

- In TDT pilot project, they adopted this model for topic tracking and first story detection.
- It associates an incoming story with topics that are already known.
- The arrival of a new story leads to a Yes/No decision for each topic.
- As a new story arrives it’s executed as a query against what’s called the inverted index.
## Vector Space Model (Contd)

### Lexicon

<table>
<thead>
<tr>
<th>TERM</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>idealist</td>
<td>71</td>
</tr>
<tr>
<td>ideal</td>
<td>803</td>
</tr>
<tr>
<td>idea</td>
<td>3586</td>
</tr>
<tr>
<td>idec</td>
<td>1</td>
</tr>
<tr>
<td>idei</td>
<td>1</td>
</tr>
<tr>
<td>idemitsu</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>identif</td>
<td>207</td>
</tr>
<tr>
<td>ideograph</td>
<td>1</td>
</tr>
</tbody>
</table>

### Postings

```
<table>
<thead>
<tr>
<th></th>
<th>(d1 : tf1) (d2 : tf2) ... (d71 : tf71)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(d1 : tf1)</td>
</tr>
<tr>
<td>1</td>
<td>(d1 : tf1)</td>
</tr>
<tr>
<td>803</td>
<td>(d1 : tf1) (d2 : tf2) ... (d803 : tf803)</td>
</tr>
<tr>
<td>1</td>
<td>(d1 : tf1)</td>
</tr>
<tr>
<td>3586</td>
<td>(d1 : tf1) (d2 : tf2) ... (d3586 : tf3586)</td>
</tr>
<tr>
<td>207</td>
<td>(d1 : tf1) (d2 : tf2) ... (d207 : tf207)</td>
</tr>
<tr>
<td>1</td>
<td>(d1 : tf1)</td>
</tr>
</tbody>
</table>
```
Vector Space Model (Contd)

- A similarity computation (cosine similarity) is executed between each of the query terms and the inverted index.
- The result of the query is documents with related terms.
- They are scored such that if they exceed a predefined threshold the story is associated to the topic.
- The threshold can be defined from series of test-run on trained data.
VECTOR SPACE MODEL (CONT'D)

- The results of this model depends strongly upon the selection of useful words and phrases from the training set.
K-NEAREST NEIGHBOR

- It is an instance based classification method.
- The system converts the incoming document into a vector and compares it to the training stories.
- Based on the cosine similarity between them, the k nearest neighbors are selected.
- The score for this document is calculated by subtracting the similarity coefficients of the negative stories from the positive ones.
- If it’s high then the story is related to the topic.
K-Nearest Neighbor

- The threshold for determining whether the score is high or not differs from a topic to another.
- So it’s hard to apply the same threshold to all topics.
**Language Model**

- W.B. Croft et al. discussed this model in their work.
- It's a statistical model for generating text.
- It estimates the probability of the query having been generated from the document model.
- The document is a good match if it is likely to generate the query.
- Experiments showed equal or less performance with vector space model.
USE OF SEMANTICAL AND CONTEXTUAL INFORMATION

- This relies on NLP (Natural Language Processing).
- As it’s difficult to differentiate between topics when they have similar terms.
- For example, two different plane crashes.
- Different approaches using this model are presented.
Lam et al. build a document representation with three vectors instead of one.

One contained concept terms, another for named entities, and the last for common terms.

Similarity of two documents was a linear combination between the three kinds of similarity.

The performance wasn’t competitively good.
USE OF SEMANTICAL AND CONTEXTUAL INFORMATION (CONT'D)

- Yang et al. they proposed two layer model.
- First the stories are classified as related to one topic, then first story detection is applied within the category.
- Named entities like persons, places,...etc were re-weighted by their effectiveness.
- This method was based on vector space model.
- It achieved good results with respect to their baseline.
SEMANTIC CLASSES

- Juha Makkonen presented a new approach based on ontology document similarity.
- He divides the term-space into semantic classes, i.e. groups of words that have meaning of the same type.
- For example: names, places, temporal, organizations, etc.
SEMANTIC CLASSES

PLACENAMES
- Mammoth_Lakes

ORGANIZATIONS

PERSONS
- Steve_Fossett

TERMS
- investigator
- plane
- crash
- millionaire
- die

TEMPORALS
- late last year
- 20080901–20081231

MAP
- Mammoth Lakes, CA
- James Steven Fossett

WHO'S WHO

GEOGRAPHICAL ONTOLOGY

TIMELINE
- 2007
- 2008
- 2009
SEMANTIC CLASSES

- He made a semantic vector out of temporal expressions.
- The term occurrences in documents are augmented with temporal references, at least the publication date.
- When comparing semantic term of two documents, the result can be increased or decreased depending on the cohesion of their temporal contexts in the documents.
- The results showed improvement with respect to the baseline.
METHODS FOR TOPIC DETECTION

- Topic detection is an unsupervised task.
- The input data is a set of topics, the output data is a certain clustering of topics.
- The type of clustering used determines whether a story can be assigned to multiple clusters or not.
- The clustering depends on the features selected in the data vector.
- This task overlaps with First Story Detection.
METHODS FOR TOPIC DETECTION (Contd)

- **K-means clustering:**
  It iterates till each story is associated with a cluster.
  If the distance between a story and the nearest cluster is below a certain threshold the story is inserted into the cluster, otherwise create a new cluster.

- **Incremental clustering:**
  Its advantage that it’s dynamic, with no restriction on the size or number of clusters.
  A drawback that the decisions can be made only once. Early mistakes can be costly.
Incremental Clustering

- Zhang et al. used the incremental clustering with some improvement to detect sub-topics as well.
- They called the approach TPIC.
- They refined the features set to improve the performance of topic detection.
- Pre-clustering operation using “age” feature of stories is used to reduce execution time of clustering process.
- The results shows it has a higher performance and less execution time than k-mean algorithms.
Hierarchical Agglomerative Clustering

- Many researches used his algorithm e.g. Dai et al. in “Event Identification within News Topics”
- It is a sequence of partitions in which each partition is nested into the next partition in the sequence.
- It is defined by disjointed clustering, which individualize each of the N documents within a cluster.
- This process is repeated in which the number of clusters decreases as the sequence progresses until a single cluster containing all N documents
HIERARCHICAL AGGLOMERATIVE CLUSTERING (Contd)

- The resulting tree of clusters is called a dendrogram, which shows how the clusters are related to each other.

Advantages

- It can produce an ordering of the objects, which may be informative for data display.
- Smaller clusters are generated, which may be helpful for discovery.
- It’s used in detecting sub-topics and relations between stories.
HIERARCHICAL AGGLOMERATIVE CLUSTERING (CONT'D)

- **Disadvantages**
  - No provision can be made for a relocation of objects that may have been 'incorrectly' grouped at an early stage. The result should be examined closely to ensure it makes sense.
  - Use of different distance metrics for measuring distances between clusters may generate different results. Performing multiple experiments and comparing the results is recommended to support the veracity of the original results.
EVALUATION METRICS

Different evaluation metrics were used like:

- Recall
- Precision
- F1 score
- Miss
- False alarm

This evaluates the results with respect to the trained data.
CHALLENGES

- So far TDT has been a very active topic of research.
- Many approaches and different models were adopted.
- Selecting the proper feature set for the data vector has been an issue, different approaches trying to solve.
- Finding a generic model for all domains has been a difficult task over the years.
GOAL

- Regarding the current events, a system is required to detect and track topics within social networks.
- Focusing on political domain, my goal is to implement a system that gives quite satisfying results about current events with all related stories using the optimal approach.
- Considering the semantic of topics and various features that help improve the results.
REFERENCES


REFERENCES

