Enhanced Temporal Random Indexing using Statistical Weighting

Presented By-
Ankush Gulati
Rohit Menon

Agenda

1. Project Description
2. Central Papers
3. Work Status
4. Implementation Details
5. Evaluation
6. Challenges Faced
7. Conclusion

Project Description

- Goal
  - Perform Temporal Random Indexing for event detection on large corpuses using statistical weights for terms.
  - Determine effects of incorporating a dimension for regional variations in semantic space model.
- Motivation
  - Improve accuracy of event detection from time ordered documents
  - Detect event trends by geographical regions

Central Papers

- Event Detection in Blogs using Temporal Random Indexing

- David Jurgens and Keith Stevens

(eETTs '09 Proceedings of the Workshop on Events in Emerging Text Types)

- Random Indexing using Statistical Weight Functions

- James Gorman and James R. Curran

(EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing)

Work Status

COMPLETED

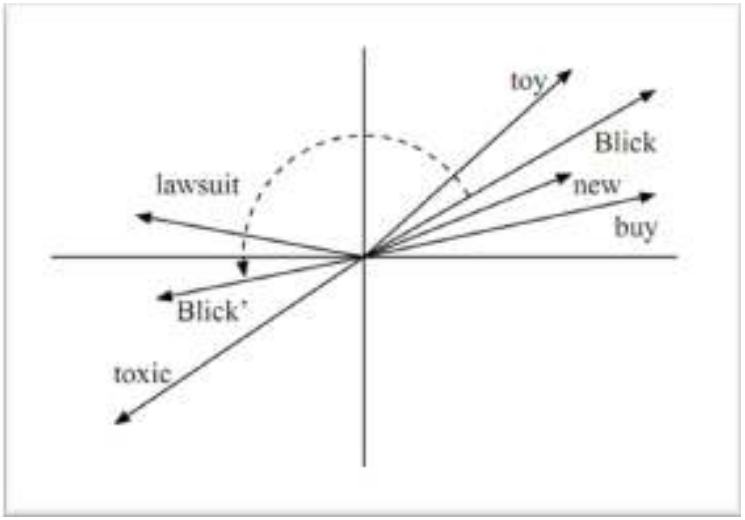
a) *Generation of Random Index Vectors per word for a corpus.*

Details:

Implemented an algorithm that generates random index vectors of 1000 dimensions for each word in the corpus.

Status:

Completed



Work Status (contd.)

COMPLETED

b) *Creating the dataset*

Details:

Implemented a website specific crawler for reuters.com that takes the search query as input and fetches all the articles associated.

Status:

Completed

(Fetched around 5000 articles from reuters.com)

1. Pass a query e.g. Osama Bin Laden
2. Download all articles about Osama Bin Laden from reuters.com
3. Classify them into different location categories.
4. Store them in <timestamp, articletext> format in per location files.

Work Status (contd.)

COMPLETED

c) *Processing & Cleaning the Corpus.*

Details:

Incorporated a functionality in the implementation that preprocesses the corpus before performing event detection mechanisms to render uniformity.

Status:

Completed

1. Replace all numbers with <num>
2. Remove all html mark-up and email addresses
3. Remove unusual punctuation, and separate all other punctuation from words
4. Remove words of 20 characters in length
5. Converting all words to lower case
6. Replacing $5 to <num> dollars
7. Discard articles with fewer than some threshold percentage of correctly spelled English words
8. Associate each entry with a numeric timestamp

Work Status (contd.)

COMPLETED

c) *Calculating Semantics for each word at a given time and location.*

Details:

Implemented the entire module that generates a Semantic Space of the corpus against time.

Status:

Completed

$$semantics(w,t) = \sum_{c_t \in d_i} \sum_{-n \leq i \leq n} index(w_i)$$

Work Status (contd.)

COMPLETED

d) *Addition of 'space' dimension to the proposed mechanism making it a corpus classified into different spaces per time.*

Details:

Implemented the entire algorithm with support for an extra dimension of space that allows user to analyze data over a period of time across different geographic locations.

Status:

Completed

$$semantics(w, t, s) = \sum_{s_i} \sum_{c_t \in d_i} \sum_{-n \leq i \leq n} index(w_i)$$

Implementation Details

a) Crawler: Creating the dataset

- Programming Language: JAVA
- Libraries Used: JSoup (Open Source HTML Parser)

b) Corpus Processor: Processing & Cleaning the Corpus

- Programming Language: JAVA
- Libraries: Stopwords lists (www.lextek.com)

c) Temporal Space Creation: Calculating Semantics for each word at a given time and location.

- Programming Language: JAVA

d) Addition of 'space' dimension to the proposed mechanism making it a corpus classified into different spaces per time.

- Programming Language: JAVA

Evaluation

a) *Computing slices per word over time and space.*

Details:

To compute a slice of word's semantics over a period of time using the semantic space already generated.

Status:

Currently working

$$slice(w) = \{(t_i, semantics(w, t_i) | w \in d_i, i = 1, k\}$$

Evaluation (contd..)

b) *Analysis of Results*

Details:

To analyze the semantic shifts of events over time at different locations for the fed corpus and see if the results are useful.

Status:

Currently working

Challenges Faced

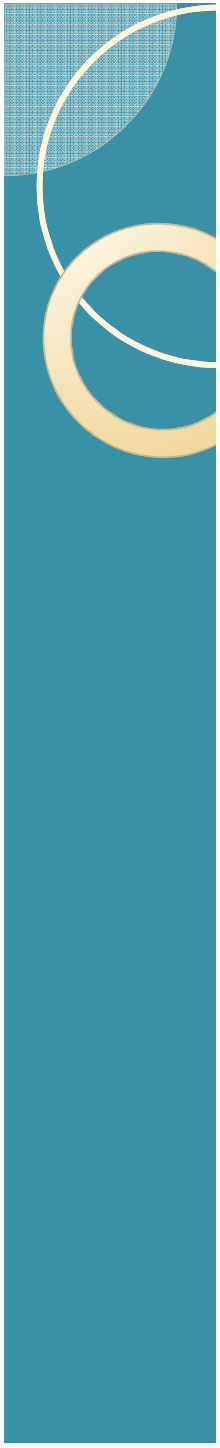
- Issues with the author's library
  - Bugs found and reported.
- Corpus Size
  - The corpus needed to test the functionality was large and very specific. Manually collecting & processing data of such size was infeasible. Thus, developed an extra functionality of dataset generator(specialized crawler).
- Handling high dimension index vectors
  - The accuracy of the implementation directly depends on the dimension size of vectors. Hence, the implementation became very complex.
- Complex data structures to handle the data sets.
  - Development of the entire project from scratch and the addition of an extra dimension required use of multilevel nested data structures which were difficult to implement & maintain.

Conclusion

- We have successfully implemented not only the proposed algorithm but also add a dimension of space to the model that will further allow evaluation of semantic shifts over time over different geographical locations.

- Furthermore, we have added the feature of removing stop words while processing the corpus.

- And lastly, we have developed a corpus generation system that collects news articles from Reuters and stores them in a processing ready format for linguistics researchers.



Thank You