Automatic Summarization Methods

State of the art

Inasse CHENNA
Malak CHERKAOUI
Hasnae DADA
Nouamane ALLOULA
Mehdi ES-SLASSI RAZZOUKI

02/03/2018

This document represents a summary of some baseline methods for microblog automatic summarization. It was done as part of the long-term project in ENSEEIHT.

Table of content

Introduction 2
1. MGraph 2
2. Temporal TF-IDF 3
3. Retweet Voting Approach 3
4. Temporal Centroid Method 3
5. Cross Media Latent Dirichlet Allocation (CMLDA) 3
6. LexRank 4
7. TextRank 4
8. SumBasic 4
9. Cluster summarizer 4
10. MEAD 5
Conclusion 5
Reference documents 5

Introduction

Automatic summarization of text has been a very important research subject during the years. Many algorithms were studied and implemented to cover this topic. Nowadays, we also talk about automatic summarization for microblogs; text summary algorithms are extended to be adapted to microblog posts which have the particularity of being short on the maximum amount of characters (140 characters for twitter for example). This document represents a brief state of the art summary about some of the existing automatic microblog summarization methods and their particularities.

1. MGraph

This summarization method aims to select a subset of images derived from a set of social media messages about an event that at the same time maximizes the relevance of the selected images and minimizes their redundancy. The framework is based on 6 major approaches:

Step 1: Filtering

A set of filters is applied in the social media items to keep only the informative ones among them.

Step 2: Multigraph generation

Creation of a multigraph that captures the similarity of items across different modalities.

Step 3: Visual de-duplication

De-duplication of messages is handled by keeping only original messages and discarding explicit reposts.

Step 4: Topic detection

To detect the topics on a main event a graph clustering algorithm is used, namely the Structural Clustering Algorithm for Networks (SCAN).

Step 5: Message selection and Ranking

Calculation of an overall importance score for each of the messages or message cliques of the filtered set, and ranks them according to it. The importance score of a message m or clique mc is a combination of two factors:

- The social attention it receives over time.
- The significance of the topic it may belong to.

Step 6: Image Ranking and Diversification

A graph-based ranking algorithm is used to diversify the top ranked social media items. Although the goal is to select a subset of images to form a visual summary, the proposed framework makes use of all the available social media items, even those who don't have any associated multimedia content.

2. Temporal TF-IDF

The algorithm is inspired by the fact that users tend to use similar words when describing a particular event, making term frequency a useful metric. The temporal TF-IDF generates summary of top terms without the need of prior knowledge of entire dataset.

The temporal TF-IDF is based on the assumption that words which occur more frequently across documents in a timeframe have a higher probability of being selected for human created multi-document summaries than words that occur less frequently.

3. Retweet Voting Approach

The number of retweets that a tweet gets in one frame time (1-hour in our case) seems at first as a useful metric to measure its ranking in a cluster. However, this method suffers from many drawbacks; The content of tweet is not always taken into consideration as many users retweet without even reading e.g. celebrities.

To overcome this problem, a normalization factor is privileged. It consists on calculating the Change of Retweet Score with time instead of the Retweet Score.

4. Temporal Centroid Method

The centroid approach takes into consideration a centrality measure of a tweet with respect to the overall topic of the cluster. The main idea behind this method is to identify posts that have high quality and most relevant to an entire cluster.

It operates in two phases:

Phase 1: It computes the cosine similarity of the TF-IDF representation of each message to its associated event cluster centroid.

Phase 2: it selects the messages with the highest similarity value

5. Cross Media Latent Dirichlet Allocation (CMLDA)

This framework's main particularity is that it handles collection of microblog posts consisting of text or both text and image. It is based on three major stages to accomplish the summarization:

Stage 1: Removal of irrelevant data

The main goal of this first step is to remove irrelevant images with a spectral filtering model based on KNN to remove irrelevant images.

Stage 2: Cross-media sub-event discovery

The goal is to discover subevents by jointly exploring the intrinsic correlation between the textual and visual aspects of microblogs and at the same time eliminating possible noisy microblog posts. To do so, two properties are explored:

- Inter-media consistency:

  Different media types of the same event should be related to certain common topics or share some common semantics. This common semantics are modeled via a subevent indicator Z,

associating one topic to each individual microblog at the opposite of normal LDA where multiple topics are associated for each document and one topic for each word.

- Intra-media discrimination:

All subevents of the same event share certain general words indicating common semantics related to the social event, while each individual subevent uniquely possesses certain specific semantics which distinguishes itself from other subevents. If the proportion of general content is large, then they may dominate the result. In order to exclude the influence of general contents and discover discriminative cues for each subevent two new latent variable R and Q are introduced. The variable indicates whether the textual or visual word is generated from the general distribution or the specific distribution corresponding to its subevent.

Stage 3: Multimedia summary generation

- Cross-Media Summarization for Microblog Texts:

Three fundamental requirements are taken into consideration to calculate the score of a post: Coverage, Significance and diversity.

- Cross-Media Summarization for Microblog Images:

Two-step approach to automatically select representative images satisfying the above two criteria:

Step 1: Partition the images within a subevent into groups via spectral clustering.

Step 2: adopt manifold ranking algorithm to rank the images.

6. LexRank

This summarizer uses a graph-based method that computes pairwise similarity between two sentences (in our case two documents i.e two posts) and makes the similarity score the weight of the edge between the two sentences. The final score of a sentence is computed based on the weights of the edges that are connected to it.

7. TextRank

This summarizer is another graph-based method that uses the PageRank [16] algorithm. This provided another graph-based summarizer that incorporates potentially more information than LexRank since it recursively changes the weights of documents. Therefore, the final score of each document is not only dependent on how it is related to immediately connected documents but also how those documents are related to other posts. TextRank incorporates the whole complexity of the graph rather than just pairwise similarities.

8. SumBasic

SumBasic uses simple word probabilities with an update function to compute the best k posts. It depends solely on the frequency of words in the original text and is conceptually very simple.

9. Cluster summarizer

This summarizer clusters the documents into k clusters using bisecting kmeans++ algorithm.

10. MEAD

MEAD is a cluster based summarizer for multi-document summarization.

Conclusion

The above algorithms can be distributed in three different categories: Frequency based summarizers, Cluster based summarizers and graph based summarizers like shown in the table below:

Algorithm	Frequency-based	Cluster-based	Graph-based
MGraph			×
CMLDA		×	
Hybrid tf-idf	×		
LexRank			×
TextRank			×
Sumbasic	×		
MEAD		×	
Cluster summarize		×	
Temporal Centroid Method		×	
Retweet	×		

Table 1: Comparative table between different algorithms

In addition to the algorithms described above, some basic baseline algorithms are also used for comparison like the Random summarizer which selects k posts randomly, or the Most Recent summarizer which selects the k most recent posts.

The above algorithms are part of what is called multi-document summarizers [5], meaning they provide a summary composed of multiple documents. There is also what is called single-document [5] summarizers that select one single post as summary like The Phrase Reinforcement Algorithm and Hybrid Document TF-IDF Summarization.

Reference documents

[1] Automatic Summarization of Real World Events Using Twitter - Nasser Alsaedi, Pete Burnap, Omer Rana

[2] Multimedia Summarization for Social Events in Microblog Stream - Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua

[3] Comparing Twitter Summarization Algorithms for Multiple Post Summaries - David Inouye* and Jugal K. Kalita+

[4] Visual Event Summarization on Social Media using Topic Modelling and Graph-based Ranking Algorithms - Manos Schinas, Symeon Papadopoulos, Yiannis Kompatsiaris

[5] Summarization of Twitter Microblogs - Beaux Sharifi*, David Inouye+ and Jugal K. Kalita*