Mining Heterogeneous Graph for Patterns and Anomalies

Ramesh Paudel, James S Park, Dr. William Eberle
Department of Computer Science

Abstract

- Knowledge discovery from disparate data sources can be very useful for gaining a better understanding of the real world.
- A better understanding of patterns and anomalies associated with a person, place, or activity, can be realized through multiple source rather than a single source.
- Graphs are a logical choice for representing real world data that have relationships and transaction.
- In this project we will investigate a novel framework capable of discovering patterns in *multiple* graphs using GBAD tool [1].

Research Objective

Our objective is not only to show that known patterns and anomalies in individual sources can still be discovered efficiently, but also that new patterns and anomalies consisting of information from multiple data sources can be identified.

Data Collection

- Dataset consist of news and tweet stream.
- News is collected using News API (newsapi.org) from 2/09/2017 to 6/23/2017.
- News API provides headlines from 70 worldwide sources and have information like author name, news title, news headline, news url, and published date.
- The news body is collective by crawling the website given by news url.
- Associated tweets 10 days before and after the corresponding news story based upon the twitter account mentioned in the body of the news is collected from Twitter REST API.
- Tweet have information like tweet date, user handle and tweet json dump.

Data Source	# of Records	Date Range
news	9,917	02/09/17 – 06/23/17
tweets	9,971	02/11/17 – 06/23/17

Table 1. Summary of heterogeneous dataset

Graph – Based Anomaly Detection

- A **labeled graph** *G = (V, E, F),* where V is the set of vertices, E is the set of edges between the vertices, and the function F assigns a label to each of the elements in V and E.
- A **subgraph** SA is anomalous in graph G if *(0 < d(SA, S) < TD)* and *(P(SA|S) < TP),* where ***P(SA|S)*** is the probability of an anomalous subgraph SA given the normative pattern S in G. **TD** bounds the maximum distance (d) an anomaly SA can be from the normative pattern S, and **TP** bounds the maximum probability of SA.
- The score of an **anomalous subgraph** SA based on the normative subgraph S in graph G is *d(SA, S) * P(SA|S),* where the smaller the score, the more anomalous the subgraph.

Data Preprocessing

- Input data was converted into *news*, *tweet* and *mixed* graph (news &tweet) using python based parser.
- These graphs reflects the relationship and transactions between various entities.
- Top five keywords from news and top three keywords from tweet are included in graph as attributes.
- Entities like place, person, organization, currency etc. are extracted from news and tweet text using Stanford Natural Language Toolkit.
- Tweet *sentiment* is calculated using Naive Bayes sentiment classifier of TextBlob tool and added on the *tweet* graph.

Graph Layout & GBAD Results

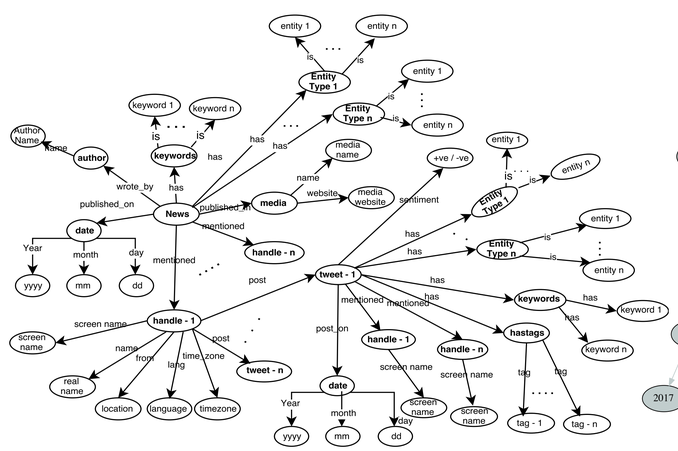

Figure 1. General Graph Layout for News and Twitter Data

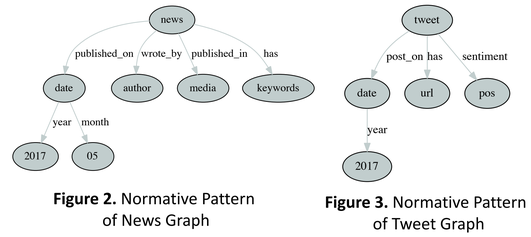

Figure 2. Normative Pattern of News Graph

Figure 3. Normative Pattern of Tweet Graph

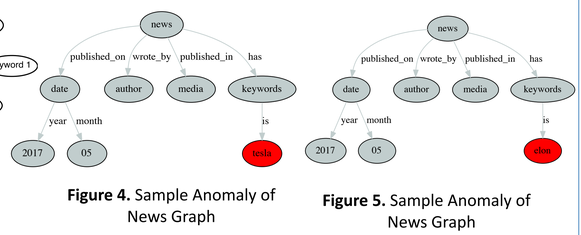

Figure 4. Sample Anomaly of News Graph

Figure 5. Sample Anomaly of News Graph

Experimentation

- Input csv data was converted into graphs using python based parser.
- Individual graphs were created for news and tweets.
- Heterogeneous graph reflecting the relationship (if the handle was mentioned in news body text) between news and tweet was also created.
- GBAD was ran on individual graphs to discover patterns and anomalies.
- Also, GBAD was ran on mixed graph being to discover patterns and anomalies.

Graph Types	# of Vertices	# of Edges
news	6,182	5,790
tweets	126,178	125,670
heterogeneous	132,360	131,968

Table 2. Summary of Graphs Used for Experiments

Discussion

- The anomalies returned by GBAD are governed by the rule, "*smaller the score , the more anomalous the subgraph*".
- The anomalies shown in Fig. 4 and 5, are reported because the keyword "elon" or "tesla" appeared few times (or just once) in whole graph resulting anomalous score to be low.
- But these anomalies doesn't convey any significant anomalies, so we would like to investigate an approach to categories keywords based on their word ontology so that "tesla" will fall under "car" or "vehicle" category and "elon" will fall under "person" or "entrepreneur" category.
- This will help us to discover normative and anomalous pattern that will better represent the more generalize information.
- Also this will help to speed up the GBAD processing time because we will have fewer unique vertices.

Conclusions

- We collected data from disparate data sources i.e. news feed and twitter feed.
- We come up with an approach to merge these data source based on their relationship and build a heterogeneous graph.
- Graph-based anomaly detection tool was used to discover patterns and anomaly on individual as well as mixed graph.
- Due to diverse news and tweet, patterns and anomalies discovered were not able to represent the real anomalies in data.
- We would like to define anomalies in terms of news and tweet stream and investigate the approach to discover them.

Future Directions

- Investigate approach for categorizing words based on their ontology to discover patterns more effectively.
- Develop and evaluate the approach of fusing heterogeneous graph streams into one.
- Ultimately use this approach for cyber crime prediction by analyzing multiple social media feeds.

References

1. Eberle, W. and Holder, L., "Anomaly detection in data represented as graphs," *Intelligent Data Analysis*, vol. 11, no. 6, pp. 663–689, 2007.

Acknowledgements

Funding provided by Tennessee Tech University, College of Engineering for achieving Carnegie classification.