Opinion Mining through NLP and Graph Database

Maralmaa Erdenebat, May @ Hnin Oo Wai

Opinion Mining through NLP and Graph Database

Maralmaa Erdenebat, May @ Hnin Oo Wai
University of Rochester

Introduction

Big data explosion took place in the early 2000 with the world annual unique data production hitting a billion gigabytes as mentioned in the study conducted by Peter Lyman and Hal R. Varian from UC Berkeley [1]. Following the rise of big data, opinion mining has become a buzz word in processing these big data for better analysis and learning. Commercial enterprises have started to realize the power of opinion mining in analyzing public sentiment about their brands. Opinion mining relies on tool such as Natural Language Processing (NLP) and Graph Databases to analyze and query text corpus. We explore the definitions and how exactly NLP and Graph Databases are used as tools.

Natural Language Processing

Though several definitions of NLP exist, the overarching concept refers to the idea of computer systems attempting to understand human languages, to analyze, interpret, or produce it and complete several tasks. These tasks could include: paraphrasing a text (input can be text, oral language or from a keyboard), translating the text to another language, providing answers to text related questions and drawing summaries or implications [2]. This is related to the NLP systems' objective to understand the true meaning and purpose of the various user's query and to produce a result that provides the intended result. There are a number of difficulties that face computer systems when implementing these tasks: lexical, structural, semantic, pragmatic and referential ambiguity [3]. Thus, NLP is rooted in disciplines in linguistics, computer and information sciences, artificial intelligence, mathematics, electrical and electronic

Opinion Mining

Opinion mining is first mentioned in a paper written by Dave et al. [6] that ideally an opinion-mining tool would "process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)." Up until now, the definition is still accurate. Sentiment analysis studies the same field of study as opinion mining and are used to broadly mean the computational treatment of opinion, sentiment, and subjectivity in text [5].

Graph Database

Graph Database is an online data management system that uses CRUD (create, read, update and delete) method on graph data model. The graph data model is composed of nodes and pointers in which nodes store the data and pointers represent relationships. It is designed so that relationships are expressed more dominantly through nodes and pointers without the need of accessing data across tables through foreign keys as in the conventional SQL databases [7]. The fast and efficient method of accessing relationships between data is significant for opinion mining which uses the relationship between words in the text corpus

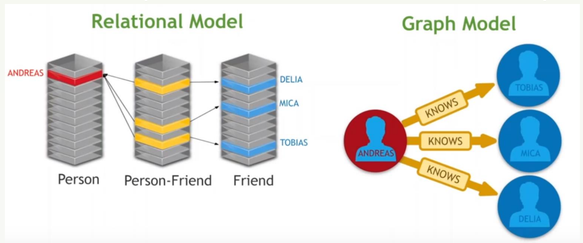


Figure 1: Difference between relation model and graph model [8]

Implementation of Graph Model in Opinion Mining

Graph Database Model could be implemented through Neo4j, an online graph database along with Cypher, the Neo4j's graph query language.

For instance, the following sentences could be added into adjacency graph through the following Cypher code: "May loves Database System. Maralmaa loves Database System. Prof. Biswas teaches Database System"

```
// Hello World!
WITH split("Prof.Biwas teaches Database System", " ") AS text
UNWIND range(0,size(text)-2) AS i
MERGE (w1:Word {name: text[i]})
    ON CREATE SET w1.count = 1 ON MATCH SET w1.count = w1.count + 1
MERGE (w2:Word {name: text[i+1]})
    ON CREATE SET w2.count = 1 ON MATCH SET w2.count = w2.count + 1
MERGE (w1)-[r:NEXT]->(w2)
    ON CREATE SET r.count = 1
    ON MATCH SET r.count = r.count + 1;
```

Figure 2: Cypher Code

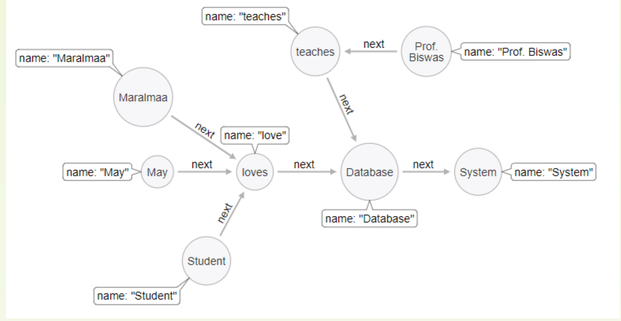


Figure 3: Resulting graph model

With a bit of modification, multiple lines of text could be processed in the similar way to generate a network of nodes representing each word and the interconnecting lines representing the relationships. The graph is made more sophisticated by including the frequency counter for each relationship and node. Thus, after setting up the graph database, we can query the frequency of the word "Database" appearing in the text corpus or find out the correlation between the two words "Prof. Biswas" and "Database" as in Figure 4. The same process could be applied to test the public sentiments on brand by analyzing tweets of a e and negative

```
// Counting the frequency of word appearances
MATCH (w:Word)
RETURN w.name AS word, w.count AS word_count
ORDER BY w.count DESC LIMIT 5
```

word	word_count
"Database"	6
"loves"	4
"System"	3
"teaches"	2
"May"	1

Figure 4: The frequency of words

Conclusion

Using Cypher in Neo4j, we were able to compute the word count of each word appearing in the text corpus. We found out the word "Database" has the highest frequency while "May" appears only one time. Through the simple implementation, the word counts of specific words were shown. For further steps in opinion mining, we could analyze whether the word is positive, neutral or negative, for instance by filtering through the word cloud provided in Google API. Through the better utilization of machine learning and larger learning datasets, a better and more accurate opinion can be mined. With graph databases, better relations can be made on multiple levels and provide deeper complexity closer to the human language.

References

[1] Lyman, Peter, and Hal R. Varian. "How Much Information?" Executive Summary, University of California at Berkeley, 18 Oct. 2000
[2] Liddy, Elizabeth D. Natural Language Processing. 2001
[3] Allen, James F. "Natural Language Processing." ACM Digital Library, John Wiley and Sons Ltd., 2003
[4] Chowdhury, Gobinda G. "Natural Language Processing." Annual Review of Information Science and Technology, Wiley-Blackwell, 31 Jan. 2005
[5] Bo Pang and Lillian Lee (2008), "Opinion Mining and Sentiment Analysis", Foundations and Trends
[6] Dave, Kushal, et al. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. 20 May 2003
[7] Robinson, Ian, et al. "Graph Databases." O'Reilly, 4 May 2015.
[8] Lyon, William. Natural Language Processing with Graphs. Natural Language Processing with Graphs, Neo4j, 18 Feb. 2016
[9] Pak Alexander, Paroubekhttps Patrick, Twitter as a Corpus for Sentiment Analysis and Opinion Mining,

Natural Language Processing (NLP)

NLP is the idea of computer systems attempting to understand human languages by analyzing, interpreting and producing it to complete certain tasks.

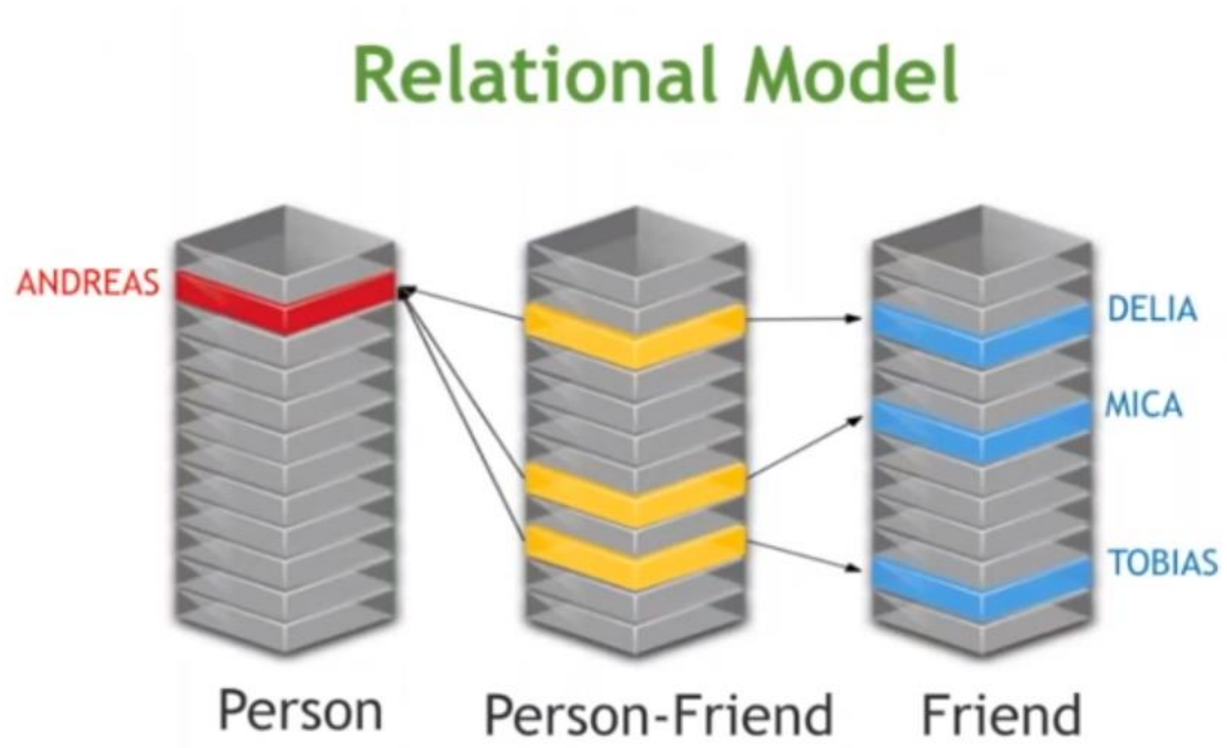
Opinion Mining

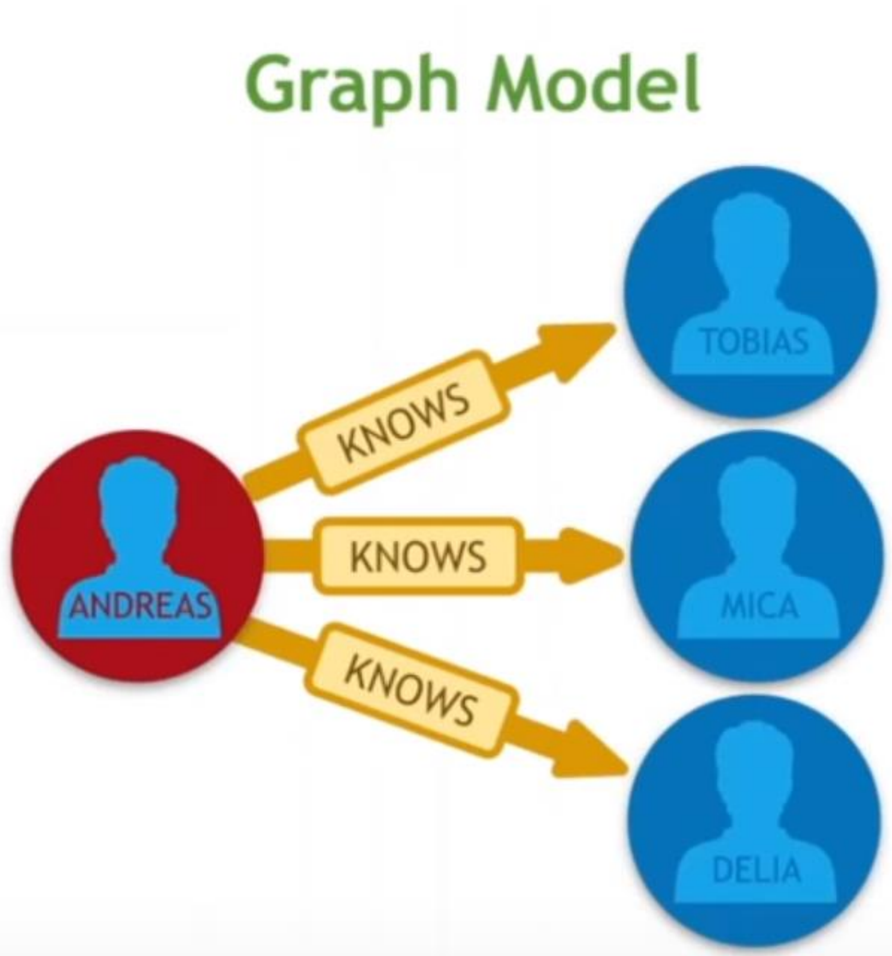
Opinion mining is a type of natural language processing for tracking the mood of the public about a particular product.

Graph Database

Graph Database is a type of NoSQL **database** that uses **graph** theory to store, map and query relationships

Relational Database vs Graph Database





Graph Database

&

Cypher