Real-Time Knowledge Management and Data-mining with Twitter

Martin Moghadam, Kalle Grafström

May 31, 2010

Abstract

Real-time tracking of human behavior on Twitter for survey, marketing and analysis.
The project focuses on Twitter for social network analysis, and the possibilities and exploits of using twitter as a data source for network analysis, exploring real-time analysis and analyzing a period of time.
Articles and research is examined and discussed, and a C# client is developed for data gathering of Twitter streams. The client gathers the data, sorts and prepares the data for knowledge management, analysis of the knowledge collected is beyond the time and scope of this project.
Twitter has proven to be a very good source for social network analysis, because of the simple structure of the data and the ease of access. The data is freely available and the are no privacy concerns. Developing a C# application is very straight forward with Web Services client model.
The application developed as a web service client could be expanded to gather data from other sources. Further development and knowledge management is discussed.

1 Introduction

Knowledge management used to identify and represent insights and experiences gained from data collected from Twitter, can prove useful in many circumstances. Some of which will be discussed in this article, others are left for further development, see discussion section.
The validity and credibility of the users identity is not discussed in detail, we simply assume the user has a valid identity, see references for further details on "The Credibility of Digital Identity Information of the Social Web".
The project uses Twitter as a data source to comprise the knowledge, the Tweets all have dates and time allowing use to develop a systems that is updated minute by minute, gathering Information in real-time. This makes it possible to examine a period of time to extract knowledge from the Tweets in that period, for instance a debate performance[1].
The project work with the data gathering for the knowledge management, creating a Twitter traffic monitoring client application.

1.1 Twitter

Twitter is a microblogging and social network service allowing users to send and read small text messages of up to 140 characters. The service has currently more than 100 million users worldwide. Users create accounts with profiles, the basic accounts are provided for free, premium account with a higher number of characters are purchasable.
The Tweets are associated to a specific user profile, the profile can have many followers that read the tweets. Each tweets has a time and date information, and can include; profile tags, topic tags and urls , this information can be used to structure and gain a greater understanding of the tweets.
The tweets contain different kinds for information as shown in figure 1.

Searching the entire web for this kind of information would yield poor results because of the wast amount of trash data that needs to be sorted though. Such trash data is almost not present in Twitter. People commonly use Twitter to share information of what they are doing and where they are. Which included the geographical information on the user.
The knowledge we are interested in, is extracted from the conversational and pass along value data. The data from Twitter is available freely, and there are currently no privacy issues with gathering data.

1.2 Premise and Usage

Twitter can and has been used to gather knowledge for:

- Real-time recommendations for topical news. [2]
- Marketing and research.
- Disaster monitoring, earthquakes, volcanoes, hurricanes. [3]

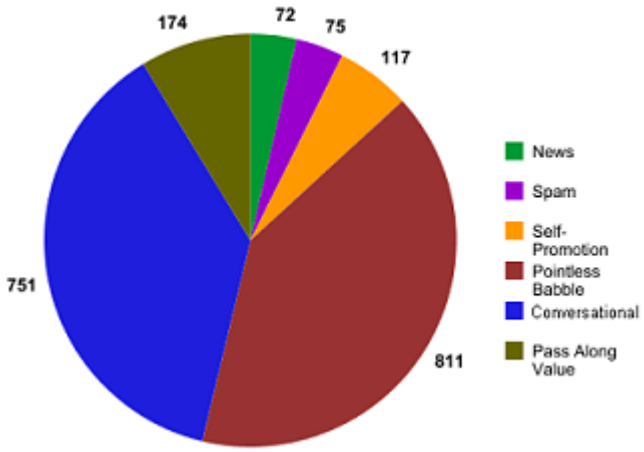

Figure 1: Content of Tweets: Source Wikipedia

- Political demonstrations. (The Iran demonstrations of 2009)
- Counter-terrorism.
- Many other possibilities.

A few people have used Twitter to make bomb treats, and encouraged assassination of politicians, these treats were not detected or monitored by an application, most of the threats were reported to the authorities by other users of the service but investigation led to no results, because the threats were not serious just pointless babble. Generally people don't announce acts of terrorism on Twitter, and there has not been any evidence of a terrorists using Twitter, making it unsuitable for counter-terrorism.
Twitter already has been used for social network analysis some of which the results can be found in the references.

2 Knowledge Management

The knowledge management discipline we work with is social networks analysis, extracting data from the social networks and examining; likes dislikes, friendships, kinship, sexual relationships, beliefs, opinions, and connections to others. The interdependencies can be represented as nodes.
These nodes can be graphically represented, the data providing a useful structured collection of knowledge.
The knowledge can provide a

2.1 Real-time Knowledge

In this case real-time means; computer system that update information at the same rate they receive information (data). The systems constantly monitors traffic, updating the data, since Twitter is a minute by minute system with time stamps, this is relatively easy. When the data is updated in real-time the knowledge can update concurrently with the data, using a multi-thread application.

3 Data Source

For the development of the application we used C# 4.0 and Visual Studio 2010, since we are already familiar with development environment and the language, it was merely a choice of convenience without much forethought.

3.1 Twitter API and development

3.2 Twitter Streams

3.3 Real-time Data

3.4 Code Example

3.5 Test and Development

No substantial test was preformed, and the development was modifying the acquired code to simplify the application and to customize it for our project.

4 Discussion

The Twitter social network analysis has great potential to monitor and gather information on human behavior and the premise we started with was validated, further development is discussed.

4.1 Further Development

Creating a more complete application of social network analysis for Twitter is possible for further development. The application could be a web app, were the user could define a period of time for analysis then the application would search and collect data from Twitter.
The application could be improved to included the following features;

- Betweenness, Closeness, Measures in social network analysis.
- Relationship detection, using the tag and link information to determine interdependencies.

- Geographic information gathering.
- Node Structuring, using the detected interdependencies to form a large node structure of the social network.
- Graphical representation, showing the interdependencies as nodes making the relationship and connection clear and easy to understand.
- Identity Credibility, create a identity credibility system to improve the accuracy of the interdependencies.
- Multi-treading and distributed solution for mass real-time Twitter social network analysis. Twitter is big and getting bigger, to analysis such a system in real-time a scalable application is needed.

4.2 Conclusion

A solution to extract data for Twitter was found and some basic analysis was test, making it a good starting point for a more complex Twitter social network analysis application. Such a project could be expanded to collection knowledge for other social networks like Facebook. Gaining another 400 million user accounts to monitor, making that a total of approximately 550 million accounts to monitor and analysis.

References

[1] Nicholas A. Diakopoulos and David A. Shamma. *Characterizing Debate Performance via Aggregated Twitter Sentiment.* ACM, 2010.

[2] Owen Phelan, Kevin McCarthy, and Barry Smyth. *Using Twitter to Recommend Real-Time Topical News.* ACM, 2009.

[3] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. *Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors.* ACM, 2010.

[4] Satyen Abrol and Latifur Khan. *TWinner: Understanding News Queries with Geo-content using Twitter.* ACM, 2010.

[5] Marc Cheong and Vincent Lee. *Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base.* ACM, 2009.

[6] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. *A Few Chirps About Twitter.* ACM, 2008.

[7] Matthew Rowe. *The Credibility of Digital Identity Information on the Social Web: A User Study.* ACM, 2010.

[8] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. *What is Twitter, a Social Network or a News Media?* ACM, 2010.

[9] Vivek K. Singh, Mingyan Gao, and Ramesh Jain. *Situation Detection and Control using Spatio-temporal Analysis of Microblogs.* ACM, 2010.

[10] Vivek K. Singh and Ramesh Jain. *Structural Analysis of the Emerging Event-web.* ACM, 2010.

[11] Bernard J. Jansen, Gerry Campbell, and Matthew Gregg. *Real Time Search User Behavior.* ACM, 2010.

[12] Amanda Lenhart and Susannah Fox. *Twitter and status updating.* ACM, 2009.

[13] *uls:*.
`http://news.yahoo.com/s/nm/20100429/wr_nm/us_venezuela_chavez`
`http://cheatedbylife.com/2010/05/08/twitter-facts-figures-infographic/`
`http://www.physorg.com/news189750438.html`
`http://theweek.com/article/index/202378/When_world_leaders_tweet`
.