COINs 2010

Sentiment Web Mining Architecture - Shahriar Movafaghi

Shahria Movafaghi\(^a\), Jack Bullock\(^b\)*

*Southern New Hampshire University, 2500 North River Road, Manchester, NH 03106-1045*

**Abstract**

In this paper we discuss the architecture of the system under development the purpose of which is to capture the sentiment of web users regarding any topic such as retail products, financial instruments (FI), or social issues like immigration. The first step is knowledge acquisition. A Sentiment Web Mining (SWM) system requires acquisition of knowledge from several sources on the web. Such knowledge may be found on blogs, social networks, email, or online news. A SWM system has customization and personalization capabilities. For our purposes, customization occurs when the SWM user can change his/her preferences to select specific sites to be used for data mining and evaluation. Personalization occurs when the system decides which sites to be used for data mining based on the user profile. The user profile dynamically changes depending on the type of user request from the system and the specific sites the user visits to verify the result of the SWM system.

The second step is knowledge storage, which involves the creation of a database. Appropriate web sites will be indexed and tagged. Taxonomy is the hardest part of this step. In this paper we will demonstrate a unique way of tagging the knowledge obtained from the web. The third step is the knowledge analysis/data mining. A SWM system will use a series of off-the-shelf knowledge analysis/data mining tools including SWM knowledge analysis/data mining engine which is based on web services technology. The type of questions used can be: 1) The volume of sentiment for a particular topic; 2) The intensity of sentiment (good or bad) for a particular topic; 3) The interrelationship between the writers of material written on the web, especially if the writer is anonymous; 4) Who is/are the leader(s) of the sentiment? If the information is maliciously posted on the web the user may want to pursue it through legal means.
The last step is dissemination of knowledge to the user(s). A SWM system uses third party visualization tools as well as web based user interfaces and reports that are written internally. The presentation component of a SWM system is decoupled from other components, namely, the process component, business rule component, and data access component for ease of maintainability.

1. Introduction

Richard Dawkins [1] described the word meme as a cultural idea (information) unit that is transferred from one person to another by non-genetic means. The quick spread of ideas can occur in social networks, emails, melodies, catch-phrases, images, and videos. Coolhunting is a term coined in the early 1990s referring to a new breed of marketing professionals, called coolhunters. It is their job to make observations and predict changes of new or existing cultural trends. The word derives from the aesthetic of "cool" [2, and 3]. There are several tools in the market to analyze the social sentiment. Reference [4] mentions several social sentiment analysis tools.

Sentiment Web Mining (SWM) systems can integrate with any third party vendor which provides adequate API (Application Program Interface). Section 3 describes the requirement and overview of the SWM system that is under development. Section 4 describes the SWM architecture. The methodology used for SWM development is described in section 3 and 4. Section 5 describes the conclusion and future work.

2. SWM System Overview

The SWM system requirements are captured through use cases [5] written using UML (Unified Modeling Language). IBM Rational Modeler [6] software is used for object-oriented analysis and design. Window 2008 R2 enterprise [7] is used for development and deployment of the system using several guest operation system and tools such as Window 2008 R2, Window 7, SQL Server 2008 R2, advance natural language processing. The virtual machine mechanism used for development is Hyper-v [8]. The virtual machine mechanism used for deployment is VMware [9].

Some of the requirements and related issues of the SWM system are listed below:
1. SWM system provides a semantic analyses capacity based on a single word or a Boolean expression. The SWM system should capture public opinion (positive, negative, and neutral) using social networking (such as Facebook, Twitter, and GossipGander), emails, news, retail sites, brand sites, images, and videos.

2. SWM provides trend analyses based on a timeline and a count of positive, negative or neutral public comments on a particular brand or subject.

3. SWM System is a web-based system with a user friendly interface.

4. SWM system provides textual and graphical reports such as pie charts, bar charts, line, and areas.

5. Using geo-tagging of websites, photographs, videos, or RSS (Really Simple Syndication), identifies the geo-position of the entity (ies). For example the SWM should be able to produce a sentiment for specific topic or product in the northeast, northwest, southeast or southwest.

6. SWM system has customization capabilities. For our purposes, customization occurs when the SWM user can change his/her preferences (web sites, positive, negative verbs, or phrases) to select specific sites to be used for different analyses. The user can also rate the comments (intensity) based on a specific web site, individual or chatting user name.

7. SWM system has personalization capabilities. Personalization occurs when the system decides which sites to be used for analyses based on the user profile. The user profile dynamically changes depending on the type of user request from the system and the specific sites the user visits to verify the result of the SWM system [10].

8. Identifies influential (leaders) from email traffic [11] or social network web sites.

9. SWM system should be able to coordinate a social network analysis with a data warehouse result for a specific business unit. For example, the SWM system provides the trend analysis between the positive sentiment and rise of a specific financial instrument.

10. SWM system would both drill down and drill up capability.
11. SWM system provides alert mechanism given specific condition for the event

12. The API (Application Program Interface) for an SWM system will make it easier to integrate with other application, especially for applications that are based on data warehousing.

13. SWM system provides web services for a series of responses such as volume of positive comments for a specific brand over a specific time line.

14. SWM system will be multi-lingual.
15. SWM system provides an XML (eXtensible Markup Language) interface.

16. SWM system is made with open architecture and the capability to integrate with other semantics analysis and data warehouse data mining tools.

There is no semantics model in the market that provides all the items listed above. Our intension is to incorporate as much as possible from third part vendors and then extend the SWM system through the addition of missing attributes using web service technology.
3. SWM System Architecture – Social Network Bus Architecture

Concurrent engineering methodology is used to develop the n-tier architecture for a SWM system [12]. Concurrent engineering in software development requires integrating different disciplines such as user interface design, software, database, and data warehouse engineering from the very start of the project.

Software systems architecture has four basic components, namely: 1) presentation component which is a system interface with the user such as a web, graphical user interface, or voice response unit. The user interface engineer works on this component. 2) process component which defines what software component (or objects in the case of object-oriented systems) to be processed by the business rule component. A software engineer works on this component. 3) business rule component processes the business rules of the organization dictated by a process component. Business rules are usually dynamic and implemented by a rule engine. Software engineers work on this component. 4) data access component retrieves, inserts, updates, or deletes data from the database and/or data warehouse. All SQL (Structure Query Language) commands are written in this layer if a relational database is chosen for the system. Database and data warehouse engineers work on this component.
engineers are also responsible for the logical and physical design of a database and/or data warehouse. Normalization techniques (usually a third normal form) is used to avoid redundancy in the database design. A dimensional model is used in a data warehouse that allows redundant data. Figure 1 shows typical software components [13].

SWM system provides the social network bus architecture for creating standards for different components of social network and third party integration. The bitmap is used for financial instrument taxonomy. The financial instrument bond maybe considered as fixed income for a group of user and mutual fund for another set of users depending on interpretation of financial instrument bitmap characteristic by individual user. In other words SWM system is flexible to allow the user to define what their interpretation of financial instrument is.

4. Conclusions and Future Research

In this paper, the requirement of a Sentiment Web Mining (SWM) system was defined. SWM system is a web based system which integrates series of sentiment analysis components. It is open architecture with many features such as volume (positive, negative, and neutral public opinion), intensity, geo-position, and trend analysis of social network web sites, emails, images and videos. SWM system identifies influential (leaders) from email traffic or social network web sites.

Object-oriented and concurrent engineering methodology is used to develop the SWM system. SWM system is n-tier architecture with clear separation of presentation, process, business rule, and data access component. It provides API, web services, and XML interface for ease of integration with other software components. Special taxonomy is used for financial instruments by using a bitmap characteristic. This taxonomy helps the users on interpretation of complex financial instrument.

Future research and work is listed below:

1. Faster retrieval of structure and unstructured data by using associative memory;

2. Identifying the percentage of correct estimates of results of SWM system and other sentiment analysis tools;

3. Develop more taxonomy algorithms for special domain such as retail, politics, and financial instruments;

4. Identify or develop more advance natural language processor using massive parallel processors; and
5. Develop and research for more advance user interface and reports, especially for financial instruments.

5. References

[4] Social Sentiment Analysis Tools
  a. www.netbase.com
  b. www.sentimentmetrics.com
  c. www.lexalytics.com
  d. www.searchenginejournal.com/alerrank
  e. www.twitratr.com