Journal of Computer Science 10 (5): 809-820, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.809.820 Published Online 10 (5) 2014 (http://www.thescipub.com/jcs.toc)

AN AGENT BASED FRAMEWORK FOR SENTIMENT CLASSIFICATION OF ONLINE REVIEWS USING ONTOLOGY

[1]P. Kalaivani and [2]K.L. Shunmuganathan

[1]Department of IT, St. Joseph's College of Engineering, Chennai, India
[2]Department of CSE, RMK Engineering College, Chennai, India

Received 2013-10-01; Revised 2013-10-05; Accepted 2014-01-15

ABSTRACT

In this study, we design and develop an agent based framework for sentiment classification of online reviews using ontology. The book review ranking is based on the sentiment classification result. We propose a novel approach with the help of the JADE platform to solve problems by non-visual automatic sentiment classification. The description of book reviews ranking are generated from the ontology based mapping. This approach employs the data extraction agent which is used to retrieve the books comments i.e., the user reviews from the specified blogs. The Second agent is the recommendation agent i.e., domain ontology is used for identifying domain related features in comments. The Third agent is feature selection agent in which XML document content is split into single sentence. Each word in the sentence is mapped with ontology. A Mapping process is used for identifying the domain related sentences in that context. These processes are used for ranking the book results based on customer reviews. The book review ranking system can be extended to other product-review easily.

Keywords: Data Mining, Sentiment Classification, Multi Agent System, JADE, Domain Ontology

1. INTRODUCTION

Web mining is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Web usage mining is a process of extracting useful information from server logs. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web content mining is the process to discover useful information from text, image, audio or video data in the web. Web content mining sometimes is called web text mining, because the text content is the most widely researched area. The technologies that are normally used in web content mining are Natural Language Processing (NLP) and Information retrieval. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site.

Sentiment classification/analysis is becoming a promising topic in the field of Customer Relationship Management (CRM). Customer profiling becomes more effective and enterprises can move towards one-to-one marketing. A basic task in sentiment classification/analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Sentiment analysis refers to the application of natural language processing, computational linguistics and text analytics to identify and extract subjective information in source materials. Sentiment analysis aims to determine the attitude of a writer with respect to some topic or the overall tonality of a document (Wanga *et al*., 2013). The attitude may be his/her judgment, affective state or the intended emotional communication.

Opinions are also important when someone wants to hear others opinions before they make a decision. There are two types of opinion: Direct opinion and

Corresponding Author: P. Kalaivani, Department of IT, St. Joseph's College of Engineering, Chennai, India

Comparisons. Direct opinions are opinion expresses on products, events, topics and people (e.g., this book is very easy to read). Comparisons express the similarities or differences between more than one object (e.g., this book explains concepts better than JAVA EDITION 5).

Consumers can use sentiment analysis to research books before making a purchase. Marketers can use this to research public opinion of their books, or to analyze customer satisfaction. Publishers can also use this to gather critical feedback about problems in newly released books.

The main objective of this study is retrieving the books name and corresponding recent reviews from specified blogs. The first agent is the data extraction agent which is used to retrieve comments about book i.e., the user reviews from the specified blogs. The second agent is the recommendation agent i.e., Domain Ontology which is used for identifying domain related features in comments. The third agent is feature selection agent in which XML document content is split into a single sentence. Each word in the sentence is mapped with Ontology. Mapping processes are used for identifying the domain related sentences in that context. These processes are used for re ranking the book results based on customer reviews.

An agent can act as an information collector, preprocessor (Othman *et al*., 2007) and classifier (Bakar *et al*., 2008) to a user.

The structure of the study is as follows. Section 2 surveys some works related to the study. Section 3 explains proposed system architecture. Section 4 shows the design of each module and implementation details. Section 5 discusses the results. Chapter 6 summarizes the study and talks about the future enhancements.

2. RELATED WORK

Several techniques were used for opinion mining tasks in history. The following few works are related to this technique.

Liu *et al*. (2012), proposed designed and developed a movie-rating and review-summarization system in a mobile environment. They used a sentiment classification approach based on Latent Semantic Analysis (LSA) to identify product features.

Tong *et al*. (2008) proposed a real time Data Mining and Multi-Agent System called DMMAS, modeling chronic disease data. They suggest that the DMMAS approach employs data partitioning and multiple agents with an option to employ heterogeneous or homogenous data mining techniques, distributing agent based processing for modeling.

Xia *et al*. (2011), in this study, ensemble framework is applied to sentiment classification tasks, with the aim of efficiently integrating different feature sets and classification algorithms to synthesize a more accurate classification procedure. The author applied, two types of feature sets for opinion mining and, three well-known text classification algorithms, namely naive Bayes, maximum entropy and support vector machines, which are employed as base-classifiers for each of the feature sets. Next, three types of ensemble methods, namely the fixed combination, weighted combination and meta-classifier combination, are evaluated for three ensemble strategies.

Zhang *et al*. (2009) presented a system for an ontology-based e-commerce product information retrieval system and proposed an ontology-based adaptation of the classical Vector Space Model with the considering the weight of product attribute. A Computer and its components related ontology has been built, which is adopted to annotate the html documents and construct concept vectors of the documents.

Mistry and Shah (2011), in this study, the author proposed an architecture for a hospital system with the help of the Jade platform. It gives an idea about different agents used and how communication occurres between them and how to manage different agents. Multi Agent System (MAS) provides an efficient way for communication between agents and is decentralized.

Wiebe *et al*. (2004) used review data from automobiles, banks, movies and travel destinations. He classified words into two classes (positive or negative) and counts the overall positive or negative score for the text. If the documents contain more positive than negative terms, it is assumed as positive document; otherwise, it is negative. These classifications are based on document and sentence level classification. These classifications are useful and improve the effectiveness of sentiment classification but cannot find what the opinion holder liked or disliked about each feature.

Zhang *et al*. (2008) used the data of customer feedback review and product review. They used Decision learning method for sentiment classification. Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Learned trees can also be re-represented as sets of if-then rules to improve human readability. These learning methods are among the most popular of inductive inference algorithms and have been

successfully applied to a broad range of tasks from learning to diagnose medical cases to learning to assess credit risk of loan applicants.

Chen and Chiu (2009) proposed a Neural Network (NN) based index which combines the advantages of machine learning techniques and semantic orientation indices to effectively classify sentiment. Tao and Tan (2004) used emotional function words instead of emotional keywords to evaluate emotional states. Hu and Liu (2004) used adjective synonym sets and antonym sets in WordNet to judge the semantic orientations of adjectives.

Existing works use semantic orientation of words for classifying positive and negative sentiments. These classifications cannot find domain related features. The proposed system introduces the combined approach of POS tagging, domain ontology and classifier intends to enhance the sentiment classification.

3. AGENT TECHNOLOGY AND MULTI AGENT SYSTEM

Agent Technology is a new concept derived from artificial intelligence. The term agent describes a software abstraction, an idea, or a concept, similar to Object Oriented Programming (OOP) terms such as methods, functions and objects. The concept of an agent provides a convenient and powerful way to describe a complex software entity that is capable of acting with a certain degree of autonomy in order to accomplish tasks on behalf of the user. But unlike objects, which are defined in terms of methods and attributes, an agent is defined in terms of its behavior. Agents itself have several characteristics that makes researchers interested to explore the agent technology. The term agent, or software agent, has found its way into a number of technologies and has been widely used, for example, in artificial intelligence, databases, operating systems and computer network literature. Therefore, an agent is autonomous, because it operates without the direct intervention of humans or others and has control over its actions and internal state. An agent is social, because it cooperates with humans or other agents in order to achieve its tasks. An agent is reactive, because it perceives its environment and responds in a timely fashion to changes that occur in the environment. An agent is proactive, because it does not simply act in response to its environment but is able to exhibit goal-directed behavior by taking initiative (Mistry and Shah, 2011).

For real world applications a single agent is not enough. So we go for multi-agents. A Multi-Agent System (MAS) is a system composed of multiple agents acting collectively to reach the goals that are difficult to achieve by an individual agent or monolithic system. In order to solve the problems mentioned above, we decided to use JADE as our implementation Tool for agents. Java Agent Development Environment (JADE) is a middleware that facilitates the development of multi-agent systems. It provides a Foundation for Intelligent Physical Agents (FIPA) compliant environment and an implementation of Multi agent system, Bellifemine *et al*. (2010).

Multi agents system is selected for the proposed system for several reasons. First of all, integrating data from various sources i.e., from various web pages is a very complex task; web pages are highly dynamic and uncertain. Secondly, agents are capable of independent action on behalf of a user or owner and can act, capture and manage information automatically when it is necessary. Thirdly, agents can interact with other external systems and can be used to manage both distributed and local knowledge. Fourthly, agents can learn from their own experience. This is particularly important in the field of web mining as the data is constantly modified and updated. This results in the system performing better over time since the agents have learnt from their previous experiences. Finally, agents have the autonomy and social ability and a multi-agent system is inherently multithreaded for control. Therefore, multi-agent approach is suitable for the development of a Sentiment Classification system.

4. METHODOLOGY

In the proposed agent based framework, the software agents are used to guide the user who has no prior knowledge in sentiment classification. The proposed system has three agents: Data extraction agent, Recommendation agent and Feature selection agent. The first agent is the data extraction agent which is used to retrieve the books comments i.e., the user reviews from the specified blogs.

The Second agent is the recommendation agent i.e., Domain Ontology is used for identifying domain related features in comments. It uses the existing ACM Classification hierarchical structure for constructing domain Ontology. Opinions are stored in XML document.

The Third agent is the feature selection agent in which the XML document content is split into a single sentence. Each word in the sentence is mapping with

Ontology. A Mapping process is used for identifying the domain related sentences in that context.

Java WordNet Interface (JWI) is used to access the WordNet database. WordNet can only recognize the following four parts of speech-NOUN, VERB, ADJECTIVE and ADVERB. Product features are usually nouns or noun phases in review sentences. We used Brill Tagger on each review to split text into sentences and to produce the part-of-speech tag for each word. Each sentence is saved in the review database along with the POS tag information of each word in the sentence.

Define the features by labeling positive or negative sentiment words. For example positive sentiment words are ‘strong’, ’clear’, ’neat’ and negative sentiment words are ‘disagree’, ’difficult’, ’bad’. The classifier classifies the sentence as either positive or negative. The final outcome of the proposed work is to re rank the book's results based on opinions of that book.

Overall system architecture is shown in **Fig. 1**. In the above architecture the focus is on domain related sentiment classification. The overall system proposes four main approaches.

4.1. Data Source

The data source proposed is www.amazon.com. This website has a lot of book reviews. These book review are downloaded using a crawler and can be used as opinions for the algorithm.

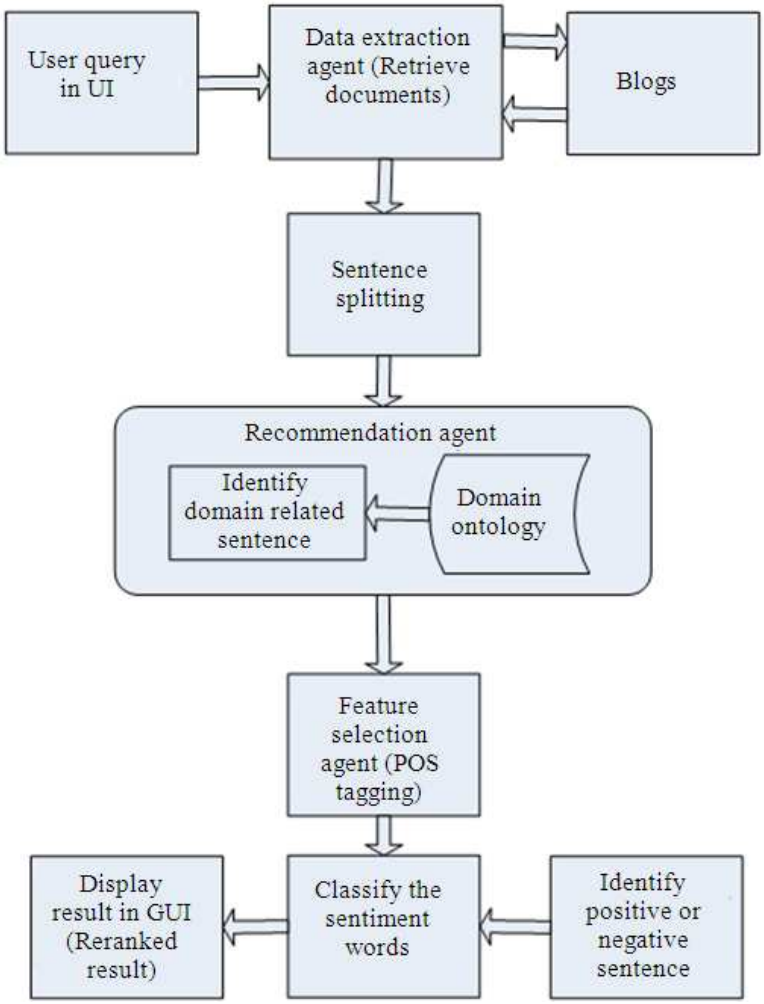

Fig. 1. Proposed system architecture

4.2. Data Extraction Agent

This is the First agent of the proposed system. Data Extraction agent is used to retrieve the book name and the corresponding book's customer reviews from specified blogs. Two different functions are used for implementing this module efficiently. The first function captures the book name and corresponding customer review links of the book from specified blogs. The second function captures the customer comments/opinion from this link. Path ascending crawling algorithm is used to implement the crawler module.

Some crawlers intend to download as many resources as possible from a particular web site. So path-ascending crawler was introduced that would ascend to every path in each URL that it intends to crawl. For example, when given a seed URL of http://www.amazon.com/Web-Haralambos/product-reviews/19339, it will attempt to crawl /Web-Haralambos/product-reviews/19339/, /Web-Haralambos/product-reviews/, /Web-Haralambos/ and /. Path-ascending crawler is very effective in finding isolated resources, or resources for which no inbound link would have been found in regular crawling. Based on this algorithm the crawler is able to capture all the book review links.

The proposed system user specifies the starting URL (www.amazon.com) and search word on web that the crawler should crawl. Data Extraction agent reads all the content and converts it into a string. The crawler captures the links only having the path containing "product-review". For example: http://www.amazon.com/Algorithms-Intelligent-Web-Haralambos-Marmanis/product-reviews/19339. After capture the links the agent should retrieves the opinions from corresponding links and store in XML document. This XML document is input for the next module.

4.2.1. Path Ascending Crawling Algorithm for Data Extraction Agent

The first function captures the book name and corresponding customer review links of the book from specified blogs. The second function captures the customer comments and opinion from this link. The user specifies the starting URL on web that the data extraction agent should crawl. Crawler reads all the content and converts it into a string:

Function 1: Retrieving book name and corresponding URL link

```
While (capture the URL)
                    {
if (URL is related to book review)
 {
 Store the URL and Book name
 }
Else continue.
Until the entire URL is captured.
                }
```

Function 2: Retrieving opinions from the link

```
While (URL list is not empty)
{
If(the URL is valid)
                    {
Read all the content and convert into string.
Remove html tags like<span>, <div><p>.
Remove an abbreviation in sentences and change into appropriate sentence.
 (Example: I've to I have)
 Store the comments in xml document.
 }
Else
List++;
 }
```

4.3. Recommendation Agent-Domain Ontology

Ontology is a formal representation of knowledge as a set of concepts within a domain and the relationships between those concepts. It is used to reason about the entities within that domain and may be used to describe the domain. A domain ontology (or domain-specific ontology) models a specific domain, or part of the world. It represents the particular meanings of terms as they apply to that domain. For example the word "ACID" has different meanings in different domains. ACID is a chemical substance in the domain of chemistry while in the domain of database management system, ACID means properties of transaction.

Domain Ontologies are used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science and enterprise bookmarking and information architecture as a form of knowledge representation.

Domain ontology is constructed that contains the domain related concepts. ACM, the world's largest educational and scientific computing society provides a hierarchical structure of Computing Systems. It uses constructing domain ontology. General Search Tree Algorithm is used for constructing the domain ontology. Normally non binary trees are used to construct the ontology. In a binary search tree, each node contains a key and points to two sub trees (Left and Right). A non binary tree contains a key and points to more than two sub trees.

In this algorithm each node of the tree contains a key and three pointers. Each node contains child, parent and brother pointer. First child only connect to parent node; all other child node of that parent connect to brother

node of previous child. This algorithm is used for inserting a new node in hierarchical order and accessing all the nodes in a fixed sequence.

4.3.1. General Search Tree Algorithm for Domain Ontology Construction

4.3.1.1. Algorithm for Node Insertion

This algorithm is used for inserting the new node in a hierarchical order.

```
Let insert node = N;
String s = Key value of Insert node.
Integer Value = Last Value of the classification.
        (For example 1.2.1 means value = 1)
If Value=1 means first child of the Parent Node then
        {
Insert Node N.Parent = address of parent node of N
        Parent.child = address of insert Node N;
        }
Else not first child of the parent node or brother of previous child node
        {
Insert Node N.Parent = address of parent node of N
                Find the brother node B;
        brother   B.brother = address of insert Node N;

        }
```

4.3.1.2. Algorithm for Traversal

This algorithm is used for accessing all the elements of the structure in a fixed sequence.

```
Node x=Root Node;
While (x! =Null)
    {
    If (x.child! =Null)
    {
    x=x.child;
    }
    If(x.child==Null&& x.brother! =Null)
        {
    Node has no child but it has a brother
        x=x.brother;                          }
    If(x.child==Null && x.brother==Null)  {
    Node has no child and brother
    x=x.parent;
    if (x.brother! =Null)
    x=x.brother;
    Else
x=x.parent;
    }   }
```

4.4. Feature Selection Agent-POS Tagger

Part of Speech Tagger (POS) is a process for marking up the words in a text as corresponding to a particular part of speech, based on both its definition as well as its context i.e., if relationship with adjacent and related words in a sentence. POS tagger module contains a set of tags such as Noun (N), Verb (V), Adjective (AJ), Adverb (AV), To (TO), Not (NOT), Conjunction (CJ), Preposition (PP), Determiner (DT) and Other (OTH). Any word from the input sentence is match with one of the tag that present in the tagset. We used the Brill Tagger algorithm for assigning tag to each word. WordNet is used for finding the POS tag of each word in the sentence. Delimiters are used to split the sentences from paragraph. The delimiters are full stop (.), expression mark (!) and question mark (?).

The Brill tagger algorithm is a method for doing part-of-speech tagging. It was described by Eric Brill. It can be summarized as an “error-driven transformation-based tagger”. It is error-driven in the sense that it recourses to supervised learning and transformation-based in the sense that a tag is assigned to each word and changed using a set of predefined rules. The output of the POS Taggers stored in an XML document.

Java API for XML Parser (JAXP) is a Java interface that provides a standard approach to parsing XML documents. JAXP provides parsers for DOM and SAX approaches to processing XML documents.

4.4.1. Brill Tagger Algorithm for POS Tagger

Generally WordNet captures only basic POS tags such as noun, verb, adjective and adverb. If the words have more than one tag (such as “book”, “saw”) brill tagger algorithm is used to find the appropriate tag.

Algorithm:

```
Known words (present in word net):
If (Probability of word is equal to one)
Assigning the tag associated to a form of the word
        Else (probability of word is less than one)
        If (The word is determiner)
                Assign the tag- DT
        Else if (The word is conjunction but not a first word in that sentence)
        Assign the Tag-CJ
        Else if (The word has more than one tag)
    Contextual rules apply for finding appropriate tag
                Else
Assigning the most frequent tag associated to a form of the word
Unknown words (out of Word net)
Assign the tag-OTH
```

4.5. Classifier

Classifier analyzes data and recognizes patterns, used for sentiment classification. It has two different types of datasets. The First dataset contains 250 positive words (such as "good", "fabulous", "recommended") and the Second dataset contains 150 Negative words (such as "bad", "not", "difficult").

All positive and negative words are ranged from 0.25 to 0.75. Classifier identifies the positive and negative words of book reviews. Equation (1 and 2) we used for calculating the cumulative positive and negative value of book review.

Positive value:

$$P = \sum_{i=1}^{n} (Wi * Ni) \quad (1)$$

Negative value:

$$N = \sum_{i=1}^{n} (Wi * Ni) \quad (2)$$

Where:

N = No. of Times a word occurs in the review
W = Weightage of the word (value from 0.25 to 0.75)
N = No of positive/Negative words in the review

Let us consider that positive values are present in $\vec{i}$ vector and negative values are present in j vector. The value of result vector $\overrightarrow{Rs}$ is $\overrightarrow{Pi} + \overrightarrow{Nj}$. Here P and N are a scalar value. Equation (1 and 2) are used for computing the value of P and N. Now this vector value's dot product with reference vector is taken the Reference vector $\overrightarrow{Rf}$ is a positive vector. Equation (3) for calculating the angle between the $\overrightarrow{Rs}$ vector and $\overrightarrow{Rf}$ vector is:

$$\theta = \cos^{-1}\left(\frac{\overrightarrow{Rs} * \overrightarrow{Rf}}{\left|\overrightarrow{Rs}\right|\left|\overrightarrow{Rf}\right|}\right) \quad (3)$$

Figure 2 shows the angle between reference vector and result vector. If the θ value is 0^o both reference book and review book are in the positive category. So the book is considered in a positive sense. If the θ value is 90° both reference book and review book are negative category. So the book is considered in a negative sense. Result vector is calculated for all other books. The angle between Reference vector and result vector calculate for remaining books. Now all the values are plotted in a graph. The book results are re-ranked based on classifier result.

4.5.1. Pseudo Code for Classifier Module

Do {
Read the XML file.
Find the Positive and Negative words of the book review
Calculate Positive value and Negative vale of each book.
Find the angle between reference vector and result vector of book using dot product.
Plot the value in graph
}
While (until the entire books are read)

5. PERFORMANCE ANALYSIS

5.1. Data Extraction Agent

The proposed system data extraction agent was used for retrieving the comments/opinions from www.amazon.com website. **Figure 3** shows the user interface of data extraction agent. User can enter the URL name (www.amazon.com) and search word in user interface.

Results are shown in user interface .Book name and corresponding book reviews are stored in XML document. Book names are stored in initial.xml file. Book reviews are stored in bookname.xml file. **Figure 4** shown as xml content of opinion about book 4.

5.2. Recommendation Agent-Domain Ontology

Domain ontology is used for identifying domain related sentences. Clas.txt is a domain ontology file that contains domain concepts. Output of data extraction agent is given as input to Recommendation Agent. **Figure 5** shows user interface of domain ontology module. User can enter the ontology file name and opinion file name in user interface.

A tree is constructed and domain related sentences are stored in domain.xml file. **Figure 6** shows a XML content of domain related sentences. Following xml contents shows domain related sentences of book 4.

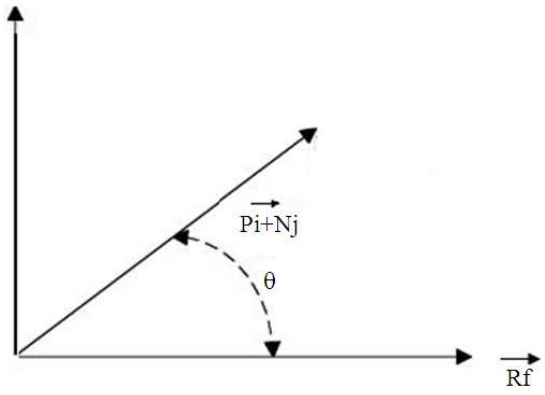

Fig. 2. Graph for Reference vector Vs Result vector

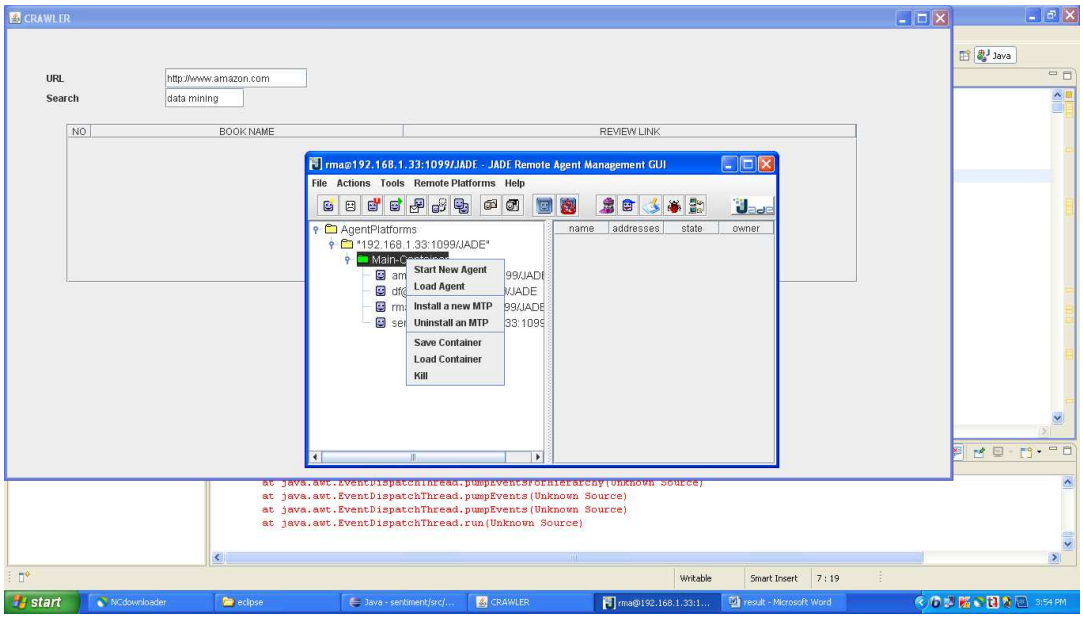

Fig. 3. User interface of crawler module

<book4>since i started a new project on behavioral targetirg in online advertising, i decided to buy linoff and berry is book: mining the web (transforming customer data into customer value). the book starts by briefly explaining data mining with a marketing point of view. the next chapter describes the three different approaches to mining the web: mining structure, mining usage patterns and mining content. in my case, i was mostly interested by mining usage patterns, for example web logs. the following chapters are very general and give an cverview of online retailing, digital content, advertising, customer value, customer tracking, etc. one good point of the book is that it contains several short industry case studies. although not technical, they help understand how things work in big industrial projects. to conclude, mining the web is an interesting introduction for people interested in applying data mining techniques to the web. altough this book answers the "what" question, it does not answer the "how". since it is a non-technical book, you will definitely need other resources for a deeper understanding of the field. the advantage is that the book is easy and pleasant to read from the beginning to the end. if you are interested by the field, you should also consult jesus mena is book: data mining your website (although a little older). i found this book to be most helpful and thorough, i was immediately inspired to practice these useful tips and easy to intuit instructions. i continue to use it for reference as a resource manual. i highly encourage anyone just getting interested in the concept of data mining, cnyone in sales, marketing, public rc ations, and analytics to start with this book first. after reading this book you will have a strong foundation into data mining applications and a vivd sense of direction on how to make it work for you personally!!!! i own berry and linoff is first two books on data mining (data mining techniques and mastering data mining); they are the best, and this book lives up to their standards. all three are great for not just teaching the technical stuff, but how and when to apply it to solve problems i really face at my company. we are just starting to look at mining the clickstream, and this is the first book i have found that cuts through the hype and really comes clean on what works and does not work. good, solid techincal information, but better is their coverage of business issues. i love all of the detailed cases. great job!</book4>

Fig. 4. Sample book review

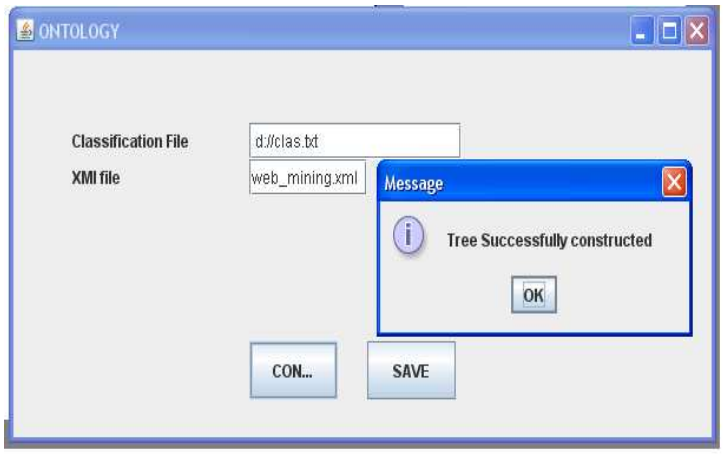

Fig. 5. User interface of Domain Ontology

<book4>the book starts by briefly explaining data mining with a marketing point of view . the following chapters are very general and give an overview of online retailing , digital content , advertising , customer value , customer tracking , etc . one good point of the book is that it contains several short industry case studies . to conclude , mining the web is an interesting introduction for people interested in applying data mining techniques to the web . altough this book answers the "what" question , it does not answer the "how" . the advantage is that the book is easy and pleasant to read from the beginning to the end . if you are interested by the field , you should also consult jesus mena is book : data mining your website (although a little older) . i found this book to be most helpful and thorough , i was immediately inspired to practice these useful tips and easy to intuit instructions . i highly encourage anyone just getting interested in the concept of data mining , anyone in sales , marketing , public relations , and analytics to start with this book first . after reading this book you will have a strong foundation into data mining applications and a vivid sense of direction on how to make it work for you personally ! i own berry and linoff is first two books on data mining (data mining techniques and mastering data mining) ; they are the best , and this book lives up to their standards . we are just starting to look at mining the clickstream , and this is the first book i have found that cuts through the hype and really comes clean on what works and does not work .</book4>

Fig. 6. Sample domain related opinions

5.3. POS Tagger

The POS Tagger module is used for marking up the word in a text as a corresponding to a particular part of speech tag based on brill tagger algorithm. Output of a domain ontology module (i.e., domain.xml) is given as inputs to the POS tagger. Tagged sentences are stored in pos.xml file. **Figure 7** shows xml content of POS tagger module. Output of POS Tagger module (i.e., pos.xml) is given as input to the classifier module. Re ranked book results store in finalresult.xml. Following xml content shows re-ranked result.

5.4. Result Comparison

Table 1 shows before and after sentiment process results. POS Tagger, Domain Ontology and classifier are the main approaches to involve the sentiment process .Re ranked results are classified by classifier based on positive and negative comments in the review.

Table 1. Sample Result comparison between before and After sentiment process

	Name of the book	
Rank	Before sentiment process	After sentiment process
1	Mining the Web: Discovering knowledge from hypertext data	Data mining the Web: Uncovering patterns in Web content, structure and usage
2	Mining the social web: Analyzing data from Facebook, Twitter, Linkedin and Other Social Media Sites	Web analytics: An hour a day
3	Web data mining: Exploring hyperlinks, contents and usage data (data-centric systems and applications)	Web mining: Applications and techniques
4	Mining the Web: Transforming customer data into customer value	Mining the Web: Transforming customer data into customer value
5	Web analytics: An hour a day	Mining the social Web: Analyzing data from facebook, Twitter, Linkedin and other social media sites
6	Data mining the web: Uncovering patterns in Web content, structure and usage	Advanced Web metrics with Google analytics, 2nd edition
7	Algorithms of the intelligent Web	Web content mining with Java
8	Web content mining with Java	Mining the Web: Discovering knowledge from hypertext data
9	Web Mining: Applications and techniques	Algorithms of the intelligent Web
10	Advanced Web metrics with Google analytics, 2nd Edition	Web analytics 2.0: The art of online accountability and science of customer centricity
11	Web analytics 2.0: The art of online accountability and science of customer centricity	semantic Web for dummies
12	Semantic Web for dummies	Web data mining: Exploring hyperlinks, contents and usage data (data-centric systems and applications)

<book4>the/DT book/N starts/V by/PP briefly/AV explaining/V data/N mining/V with/PP a/N marketing/N point/V of/PP view/V . the/DT following/N chapters/N are/V very/AV general/POSITIVE and/CJ give/V an/DT overview/N of/PP online/OTH retailing/V , digital/AJ content/V , advertising/V , customer/AJ value/N , customer/AJ tracking/N , etc/OTH . one/AJ good/POSITIVE point/N of/PP the/DT book/N is/V that/CJ it/N contains/V several/AJ short/AJ industry/N case/V studies/V . to/TO conclude/V , mining/V the/DT web/N is/V an/DT interesting/POSITIVE introduction/N for/CJ people/V interested/V in/PP applying/V data/N mining/V techniques/N to/TO the/DT web/N . altough/OTH this/DT book/N answers/V the/DT "what"/OTH question/V , it/N does/V not/NOT answer/V the/DT "how"/OTH . the/DT advantage/N is/V that/CJ the/DT book/N is/V easy/POSITIVE and/CJ pleasant/AJ to/TO read/V from/PP the/DT beginning/N to/TO the/DT end/N . if/CJ you/N are/V interested/V by/PP the/DT field/N , you/N should/MV also/AV consult/V jesus/N mena/OTH is/V book/V : data/N mining/V your/DT website/N (although/OTH a/N little/POSITIVE older)/OTH . i/N found/V this/DT book/N to/TO be/V most/AV helpful/POSITIVE and/CJ thorough/AJ , i/N was/V immediately/AV inspired/V to/TO practice/V these/DT useful/POSITIVE tips/N and/CJ easy/POSITIVE to/TO intuit/V instructions/N . i/N highly/AV encourage/POSITIVE anyone/OTH just/AV getting/N interested/V in/PP the/DT concept/N of/PP data/N mining/V , anyone/OTH in/PP sales/N , marketing/V , public/AJ relations/N , and/CJ analytics/OTH to/TO start/V with/PP this/DT book/N first/AJ . after/CJ reading/N this/DT book/N you/N will/V have/V a/N strong/AJ foundation/N into/PP data/N mining/V applications/N and/CJ a/N vivid/AJ sense/N of/PP direction/N on/PP how/CJ to/TO make/V it/N work/V for/CJ you/N personally/AV ! i/N own/AJ berry/N and/CJ linoff/OTH is/V first/AJ two/AJ books/N on/PP data/N mining/V (data/OTH mining/V techniques/N and/CJ mastering/V data/N mining)/OTH ; they/N are/V the/DT best/POSITIVE , and/CJ this/DT book/N lives/V up/PP to/TO their/DT standards/N . we/N are/V just/AV starting/V to/TO look/V at/PP mining/V the/DT clickstream/OTH , and/CJ this/DT is/V the/DT first/AJ book/N i/N have/V found/V that/CJ cuts/N through/AV the/DT hype/N and/CJ really/AV comes/V clean/POSITIVE on/PP what/CJ works/V and/CJ does/V not/NOT work/V .</book4>

Fig. 7. Sample output of POS tagger module

5.5. Performance Measure

Accuracy and precision are important measures for evaluating the sentiment classification system. The following formula is used for calculating the accuracy and precision of the system Equation (4 and 5):

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (4)$$

$$\text{Precision} = \frac{TP)}{(TP + FP)} \quad (5)$$

Where:
TP = True Positive
TN = True Negative
FP = False positive
FN = False negative

True positive means number of positive sentences which the system predicted as correct. False positive means number of positive sentences which the system predicted as wrong. True negative means number of negative sentences which system predicted as correct. False positive means number of negative sentences which system predicted as wrong.

5.5.1. Accuracy

Figure 8 illustrates the accuracy of the sentiment process. A Line graph is drawn between percentage of accuracy and book opinion. Percentage of accuracy is plotted in Y axis and books are plotted in X axis. Graph explains that the number of inputs increases performance of classifier also improves.

5.5.2. Precision

Figure 9 illustrates the precision of the sentiment process. Line graph is drawn between percentage of precision and book opinion. Percentage of precision is plotted in Y axis and books are plotted in X axis.

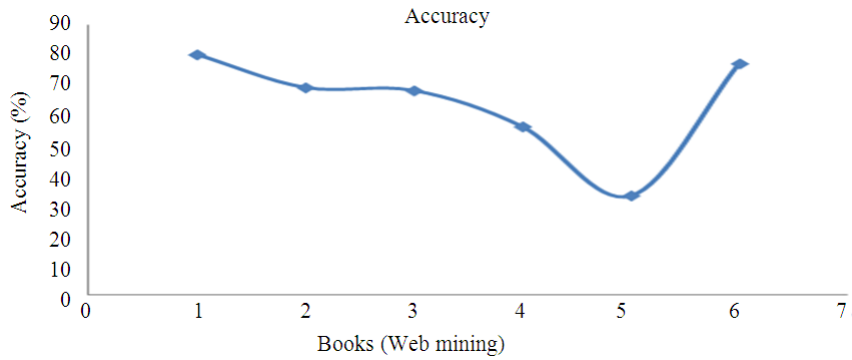

Fig. 8. Accuracy

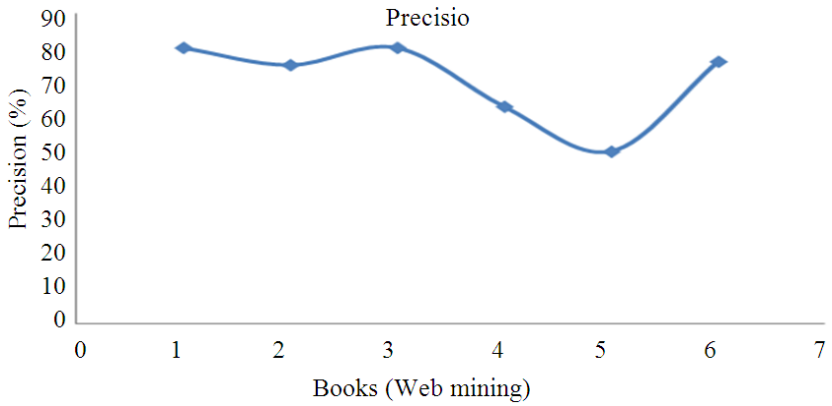

Fig. 9. Precision

6. CONCLUSION

Sentiment classification of reviews is an important objective and challenge in customer relationship management. The proposed system uses the online data source (www.amazon.com) for implementing the study. The system uses agents to classify the user opinions. In this study, agent retrieved the books name and corresponding recent reviews from specified blogs. The first agent is the data extraction agent which is used to retrieve the books comments i.e., the user reviews from the specified blogs. The Second agent is the recommendation agent i.e., Domain Ontology is used for identifying domain related features in comments. The Third agent is feature selection agent in which XML document content is split into a single sentence. Each word in the sentence is mapped with Ontology. A Mapping process is used for identifying the domain related sentences in that context. These processes are used for re ranking the book results based on customer reviews.

Moreover, we used only 500 sentiment words to evaluate ontology based sentiment classification. More sentiment words need improve the classifier. Thus it will be our future work to achieve greater accuracy.

The proposed method can also be applied to other languages. A multilingual sentiment-based lexicon needs to be developed in the future. The proposed system used single domain ontology for identifying domain related sentences. Further research can also used multi domain ontology for identifying domain related sentences.

7. REFERENCES

Bakar, A.A., Z.A. Othman, A.R. Hamdan, R.I. Yusof and R. Ismail, 2008. Agent based data classification approach for data mining. Proceedings of the International Symposium on Information Technology, Aug. 26-28, IEEE Xplore Press, Kuala Lumpur, Malaysia, pp: 1-6. DOI: 10.1109/ITSIM.2008.4631677

Bellifemine, F., G. Caire and T. Trucco, 2010. Jade Programmer's Guide. University of Parma.

Chen, L.S. and H.J. Chiu, 2009. Developing a neural network based index for sentiment classification. Proceedings of the International Multi Conference of Engineers and Computer Scientists, (ECS' 09), International Association of Engineers, Hong Kong, pp: 744-749.

Hu, M. and B. Liu, 2004. Mining and summarizing customer reviews. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 22-25, ACM Press, New York, USA., pp: 755760. DOI: 10.1145/1014052.1014073

Liu, C.L., W.H. Hsaio, C.H. Lee, G.C. Lu and E. Jou, 2012. Movie rating and review summarization in mobile environment. IEEE Trans. Syst. Man Cybernet., 42: 397-407. DOI: 10.1109/TSMCC.2011.2136334

Mistry, M. and D. Shah, 2011. Impact of multi-agents in hospital environment. Proceedings of the International Conference on Intelligent Systems and Data Processing, (ISD' 11), pp: 32-36.

Othman, Z.A., A.A. Bakar, A.R. Hamdan, K. Omar and N.L.M. Shuib, 2007. Agent based preprocessing. Proceedings of the International Conference on Intelligent and Advanced Systems, Nov. 25-28, IEEE Xplore Press, Kuala Lumpur, pp: 219-223. DOI: 10.1109/ICIAS.2007.4658378

Tao, J. and T. Tan, 2004. Emotional Chinese talking head system. Proceedings of the 6th International Conference on Multimodal Interfaces, Oct. 13-15, ACM Press, New York, USA., pp: 273-280. DOI: 10.1145/1027933.1027978

Tong, C., D. Sharma and F. Shadabi, 2008. A multi-agents approach to knowledge discovery. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Dec. 9-12, IEEE Xplore Press, pp: 571-574. DOI: 10.1109/WIIAT.2008.418

Wanga, S., D. Li, L. Zhao and J. Zhang, 2013. Sample cutting method for imbalanced text sentiment classification based on BRC. Knowl. Based Syst., 37: 451-461. DOI: 10.1016/j.knosys.2012.09.003

Wiebe, J., T. Wilson, R. Bruce, M. Bell and M. Martin, 2004. Learning subjective language. Comput. Linguist., 30: 277-308. DOI: 10.1162/0891201041850885

Xia, R., C. Zong and S. Li, 2011. Ensemble of feature sets and classification algorithms for sentiment classification. Inform. Sci., 181: 1138-1152.

Zhang, C., W. Zuo, T. Peng and F. He, 2008. Sentiment classification for Chinese reviews using machine learning methods based on string kernel. Proceedings of the 3rd International Conference on Convergence and Hybrid Information Technology, Nov. 11-12, IEEE Xplore Press, Busan, pp: 909-914. DOI: 10.1109/ICCIT.2008.51

Zhang, L., M. Zhu and W. Huang, 2009. A framework for an ontology-based e-commerce product information retrieval system. J. Comput., 4: 436-443. DOI: 10.4304/jcp.4.6.436-443