ABSTRACT BOOK

M.TECH. DISSERTATION
(JAN-MAY 2017)

Department of Computer Engineering

YMCA UNIVERSITY OF SCIENCE AND TECHNOLOGY,
FARIDABAD-121006

2017

INDEX

19	606	Divya Sharma	context extraction for followees Recommendation in Twitter	Dr. Jyoti	23
20	607	Monika Choudhary	Implementation of security measure in Cognitive Radio network against object function Attack	Dr. Preeti Sethi/Ms. Poonam	24
21	608	Neha Gupta	Design of efficient Routing protocol using power awareness in ADHOC network	Dr. Parul Tomar	25
22	609	Nisha Jha	Automatic Testing of web applicaions using Selenium and Jmeter	Dr. Rashmi Popli	27
23	610	Ojasvini Dhingra	Click Prediction using Data Analaytics	Dr. Rashmi Popli	28
24	611	Pooja Rani	Explicit Trust calculation in online social network	Dr. Sapna Gambhir	29
25	612	Sadia Parveen	Parallel frequent itemset mining using interval intersaction and Bit Vector Approach.	Dr. Neelam Duhan	31
26	613	Shefali Raina	Sentiment Analysis & summarization of reviews in food domain	Dr. Komal Bhatia	32
27	614	Suruchi Garg	A technique for review summarization based on sentiment Analysis	Dr. Komal Bhatia	34
28	615	Tamanna Sachdeva	A novel architecture for opinion mining using star ratings	Ms. Shruti Sharma	35
29	617	Vrinda	A novel approch for sentiment analysis using deep learning	Dr. Komal Bhatia	36
30	618	Yashasvee Shukla	Travel based recommendation system based on collaborative filtering	Dr. Jyoti	37
31	MNW-641-2k15	Arif Khan	Smart document clustering (To remove document duplicacy)using TF-IDF & cosine similarity approch	Ms. Deepika	38
32	642	Anjali	Content extraction from social media	Dr. Sonali Gupta	39
33	643	Deeksha	A novel approach for PCA based summarization	Dr.Komal Kumar Bhatia	40
34	644	Deepti	Secure cloud based data dissemination in VANETs	Dr. Sapna Gambhir	41
35	646	Divya	Domain specific feature extrqaction for efficient opinion mining	Ms. Shruti/ Ms.Payal Gulati	42
36	647	Jyoti	Hybrid approch for effieient opinion mining	Dr. Naresh Chauhan /Dr. Payal Gulati	43
37	648	Masoom	Secure data aggregation in wireless sensor networks	Dr. Sapna Gambhir	44
38	649	Preeti	Improved focused crawler using graphical property,web page classifier and link evaluation	Dr. Ashutosh Dixit	45

39	650	Reena	Review spam Detection combined approch of behavioral analysis and structural analysis	Dr. Naresh Chauhan /Dr. Payal Gulati	46
40	652	Shivam	A study and development of gesture control mechanism for handicapped using MATLAB	Mr. Umesh Kumar	47
41	653	Srishti	A novel approch for novelty detection via to pic modeling	Mr. Sushil	48
42	654	Shaveta	ontology based extraction using web mining	Mrs. Deepika	49
43	656	Sonam	Semantic Similarity between two Documents using topic MAP	Mrs. Mamta Kathuria	50
44	657	Sudarshan	Design of MD5 based authentication protocol using mobile agent	Mr. Umesh Kumar	51
45	658	Yashasvi	Relevant content Extraction & Text Summarization	Dr. Ashutosh Dixit	52

Semantic Similarity between Documents Using Ontology Tree View

There is a lot of information present on web. Some documents belong to same group and others to different group. But it is not easy to check if more than one document belongs to same group or different group if they do not show some similarity. Depending on methods we get the results, but sometimes we do not satisfy with results. There are some methods which are normally used to check the similarity between documents such as Keyword matching method, semantic similarity method, ontology similarity method. We can use keyword matching if exact keywords present in both documents and semantic similarity provides better result in case of semantic words, and ontology similarity method works on relationship. But sometimes all methods do not provide better results in some specific conditions. of If there are not exact keywords present in documents then keyword matching do not provide efficient results. If there will be a choice of semantic similarity which can solve the problem of exact keyword matching along with semantic words checking. But in case of unavailability of both conditions what will be the choice. But Both documents contain a real life existing relationship without having exact keywords along with semantic words. Now there is a need of hybrid method which can solve this problem. This problem can be solve by using semantic similarity based on ontology which can serve both purpose. The computation of semantic similarity between words is important in information retrieval, knowledge acquisition and many other fields. This semantic similarity calculation starts from ontology tree view with additional attributes addition and then there is an application of semantic check at each root. This work introduces Semantic similarity method using Ontology which is the hybrid of all these methods for providing better results. This work shows comparisons in the results of all methods also.

Anju Rani

CE-621-2K15

Test Case Prioritization using Clustering

Software testing is the most important part of software development life cycle. There are various types in software testing which have their own different functionalities. Among them regression testing is most useful functional type of testing which is done in the software maintenance phase. This testing is used to check the errors when any change is made in the existing system. To make system efficient and effective, techniques of test case prioritization are used. The reduction in the cost of testing and fault detection capabilities of testing should be done by test case prioritization. This technique is also applied on different algorithms to improve their efficiency. Many clustering algorithms may also use test case prioritization method to increase the efficiency of software. Prioritization techniques that incorporate a clustering approach and utilize code coverage, code complexity to increase the effectiveness of the prioritization. Clustering is the process of partitioning a group of data points into a small number of clusters. The objective of clustering is to partition a set of objects into clusters such that objects within a group are more similar to one another than patterns in different clusters. Our main objective of thesis is to reduce the execution time by yielding maximum faults of regression testing.

This thesis work presents an approach to prioritize regression test cases based on three factors. These factor are clustering, rate of fault detection, percentage of fault detected. The proposed is compared with prioritization techniques using APFD metric that is used to find the degree of faults detected. The prioritized test cases yield better fault detection than the non – prioritized test cases.

Annu Mohar

MCE-622-2K15

Sentimental Analysis on Opinion Data Summarization using Clustering & Support Vector Model

User behavior analysis could be a method throughout that the polarity (i.e. positive, negative or neutral) of a given text is set, generally objective or unit which approaches to deal with this problem, particularly, machine learning approach or lexicon primarily based approach, in this scheme using text summarization using document clustering we will analyze the sentiments of the query on the context. It is very difficult for human beings to manually summarize large documents of text and then analyze the sentiments over the same. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An abstractive summarization method consists of understanding the original text and re-telling it in fewer words. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

Anu Sharma

MCE-623-2K15

Discovery of Entity Synonyms for Efficient Web Search

The current internet search is based on finding entities in contrast to previous approaches that are based on query string matching. The web queries have become more and more pin-pointed aiming to find values relating to specific entity in a specific context of place, time, etc. For example, information pertaining to a movie-show, a particular train, news paper of a particular date, performance of a particular stock etc. Entities are exceptionally powerful component for search. So, the use of entities in serving peoples' daily information needs and its synonyms is increasing day by day.

One major hurdle in entity understanding is that the same entity is referred by variety of alternate names, i.e. entity synonyms or entity references. These references vary with the heterogeneous contexts of the web and one may not be getting the required answers to his/her query because of these varied entity references known as entity synonyms. Existing approaches that are used to discover entity synonyms have limited coverage and diversity. These entity synonyms cannot be handled through lexical resources like Wordnet. Search engine usually ranks pages based upon how terms from searcher's query appear on those pages and if relevant words are not used in their search then they may miss the pages and the information relevant to them. This problem becomes more serious as the number of terms in the query increases. Therefore, every search engine will have to create its own mechanism for finding the entity synonyms of a particular entity in order to properly answer the users' query.

So, the proposed work is a recent trend that avoids expensive human laboring and increases coverage and diversity of synonyms by the use web search engine. It presents a dynamic approach that generates rich-set of entity synonyms using inbound anchor text, context and trailing part of URL. The Proposed method is verified with experiments on real-life data sets showing the significance it brings in improving user search experience. Moreover, the coverage of web queries is increased with good precision.

Anuradha Singh

MCE-624-2K15

Wireless Sensor Network Security Attacks

Wireless sensor networks (WSNs) are an area of great interest to both domain and business. WSNs raise a large range of military, industrial, scientific, civilian and business applications to a different level, permitting, value effective sensing, particularly wherever human observation or ancient sensors would be undesirable, inefficient, expensive, or dangerous.

Even from their earliest applications, sensor networks are targeted for attack by adversaries with an interest in intercepting the information being sent, or in reducing the flexibility of the network to hold out its mission. There are lots of potential attacks against WSNs, that have totally different objectives, are performed at totally different levels, and result in totally different consequences.

Wireless sensors have limited energy and procedure capabilities, creating several ancient security methodologies tough or not possible to be utilized. Also, they're usually deployed in open unattended areas, allowing physical attacks like jamming, node capture and tampering.

This project designed and developed a new protocol that stops wormhole attacks on wireless networks.

Thus, before making an attempt to style a secure communication mechanism, it's important to properly study and understand the WSN design, the hardware and code specifications of the sensor nodes, the preparation surroundings, the production price, the benefits and mostly the constraints, that might influence the WSN style. Particular stress is given to the security demand of authentication. Before allowing a node to access real-time information from another sensor node, authentication should be ensured. The projected scheme provides mutual authentication of the sensor nodes, with the utilization of sinusoidal functions that operate as pseudorandom range generators, keeping in mind the nodes' restricted computational, power and communication capabilities.

Deepak Kumar

MCE-625-2K15

Opinion Mining with Spam Detection using Ontology

When purchasing a product for the first time one usually needs to choose among several products with similar characteristics. Companies use to promote their brands and products pointing out good characteristics avoiding to mention the poor ones. The best way to choose the most suitable product is to rely upon the opinions of others. Analysis of such texts is a more productive way of collecting user information than the traditional structured data collection by surveys where people are usually unwilling to take the time to answer presupposed questions. Conversely, sentiment analysis "listens" to published opinions and answers "the unknown". Sentiment analysis or opinion mining is the computational study of people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics and their attributes. This is an Information Extraction task which is technically very challenging but also practically very useful. In the current scenario, the data on the web is growing exponentially. Social media is generating a large amount of data such as reviews, comments, and customers' opinions on a daily basis. This huge amount of user generated data is worthless unless some mining operations are applied to it. As there are a number of fake reviews so opinion mining technique should incorporate Spam detection to produce a genuine opinion. Nowadays, there are a number of people using social media opinions to create their call on shopping for product or service. Opinion Spam detection is an exhausting and hard problem as there are many faux or fake reviews that have been created by organizations or by the people for various purposes.

In this work many approaches are used to yield better results. The techniques include ontology, Geo location and IP address tracking, Spam words Dictionary using Naïve Bayes, Brand only review detection, tracking account used. This thesis successfully addresses the significant challenges and has resolved many of them. The main focus of the thesis was to find such a technique for analyzing sentiments posted on web that can efficiently perform sentiment analysis with spam detection in a fast and accurate manner.

Divya

MCE-627-2K15

Predictive Capability Enhancement using Open Source Dataset

With the development and penetration of data mining within different fields and industries, many data mining algorithms have emerged. The selection of a good data mining algorithm to obtain the best result on a particular data set has become very important. What works well for a particular data set may not work well on another. The goal of this thesis is to evaluate the classification algorithms and characteristics of data sets by first building a file of data sets, their characteristics and the performance of a number of algorithms on each data set; and second applying supervised classification analysis to this file to find the classification accuracy.

Five classification algorithms were applied to 79 data sets. The five classification algorithms used were BayesNet, Logistic Regression, Kstar, Decision Table, and Logistic Model Tree (LMT).

The major discovery made by this thesis is that performance of classification techniques varies with different data sets. Factors that affect the classifier's performance are: 1. Data set, 2. Number of instances and attributes, 3. Type of attributes, 4. System configuration.

Experiments performed for this thesis also allowed the comparison of the performance of the 6 classification algorithms with their default parameter settings. It was discovered that in terms of performance, the top three algorithms were 1.LMT, 2 Logistic Regression, 3. BayesNet and other two algorithms Kstar and Decision table did not perform well.

Jyoti

MCE-628-2K15

Sentiment Analysis with Sarcasm Detection

Internet is a source of huge information. By just one click on your laptop or mobile phone you can gather information about anything in the world. E-commerce, social networking, education, entertainment, online tourism etc are widely used web applications. These application providers allow users to share their feedbacks, reviews and suggestions online on their websites. These reviews are used by other people for decision making when buying a new product or services and by organization itself to offer better services to their customers. Sentiment analysis is a Natural language processing or Information Extraction task that process the user's views or opinions and identifies whether the review is positive, negative or neutral.

Sarcasm is most common challenge faced during sentiment analysis. In sarcastic review basically a positive word is used to express a negative opinion about an object or service. For handling sarcasm real world knowledge is required. This thesis presents domain specific sentiment analysis using supervised learning technique along with sarcasm detection, where airlines services are chosen as domain. The system is able to classifying the reviews using Naive Bayesian machine learning technique and give positive or negative result about an airline's reviews. Method to detect sarcasm is applied on Naïve Bayesian output. The system is able to handle sarcastic reviews and give more accurate result. In the process of constructing the system, many previous works are reviewed and some of them are applied to the thesis and different methods are compared for reaching a best solution.

The thesis is mainly developed to provide best quality of service to the customers. An organization is interested in customer's perception about its services. This information is used to improve the services and new marketing strategies. Other people can use this piece of information to choose best option from the available ones based on their requirements. Organization can use customer's suggestions to add on them in near future to maximize the profit. Customer can use these reviews to make a decision based on their requirements.

Laxmi Yadav

MCE-629-2K15

Event Extraction from Twitter

With the fast growth of social media, interest is increasing in detecting popular events from tweets. Event extraction is a work of identifying events from tweets or database of tweets. Each and Every day, hundreds of Megabytes of current stories are being added into the news archives of the major news agencies, containing much important and interesting news. The application of events includes time line formation, text summarization, FAQs etc. Events are also defined as predicates (statements) that describe the circumstances in which something holds true. Events may be expressed by means of verbs, adjectives, predicative clauses, or prepositional phrases. Event extraction refers to the task of discovering and saving structured representations of major life events from tweets with related attributes and properties, which are often, categorized by complex, and nested argument structures involving multiple entities. It is one of the atomic operations in detection and involvement among entities and other information in tweets or documents. Many of previous research on event extraction have focused on textual level extraction such as News articles, text summarization and Blogs, whereas few examples can be found on event extraction from noisy text such as tweets. For instance, tweets are short and self-contained which make them lack of useful information such as contextual information.

The target of this research is to develop algorithm and methodology that extract and efficiently conclude major life events, the so-called (breaking news), extracted from social media. This task is useful for the professional journalists; it helps them to utilize social media as an information source helps to get a handle on with the lot of information. Twitter texts exclusive is its word count limitation which causes extensive usage of acronyms and other abbreviations. Event extraction combines knowledge and experience from a number of domains; it includes computer science, linguistics, data mining, and artificial intelligence.

Monika Gupta

MCE-630-2K15

A Study on Detection of Misbehaving nodes in Ad Hoc Routing

Adhoc network is a collection of wireless networks. Each node communicate in a radio communication range. A MANET is a group of multi-hop wireless adhoc networks. Each node transfer the message to the other node through the wireless networks. This network is not fully secured so there are more chances of attacker .The success of mobile adhoc network depends on people's confidence in its security.

To identify the malicious node different techniques are used. The work in the thesis deals with the malicious node for secure data transmissions.

Monika Jatiwal

MCE-631-2K15

Design and Analysis of Schedules on cloud computing platform with Linear Regression Based Ordinal Optimization

In the current scenario, constant change in the user requirements and increase in demand for large scale computational resources and high infrastructure has led to the development of cloud computing. Cloud Computing involves the allocation of these resources to the end user in a manner such that these resources get utilized effectively and efficiently.

Task scheduling problem in cloud computing is NP-hard problem. If there are m resources and n tasks available, so there are many ways of scheduling n tasks to m resources. We can optimize this problem to achieve goals and objectives in cloud computing environment. Depending upon user requirements, tasks with high computational granularity results in composition of scientific workflows in cloud computing environment.

In Cloud computing distributed resources are used on demand basis without having the physical infrastructure at the client end. Cloud has a large number of users and has to deal with large number of task, so scheduling in cloud plays a vital role for task execution. Scheduling of various multitask jobs on clouds is considered as an NP-hard problem. In order to reduce the large scheduling search space an ordinal optimization (OO) method has already proposed. The proposed work is divided in two parts.

The first part consists of designing candidate schedules which will act as test bed for applying ordinal optimization. Then using appropriate selecting rules for selecting Good Enough set, G and accepted set S. Finally GΠS is find which will give optimum schedules.

In second part different Loads (No. of Cloudlets) are applied on these optimum schedules. In order get the equation for minimum Makespan for optimum schedule for a given load on the cloud Linear Regression technique is applied. So this best fitted equation will give the minimum Makespan for the given load. In this work CloudSim version 3.0 is used to test and analyse the policies.

Monika Yadav

MCE-632-2K15

Data Security in Cloud Computing using Steganography and AES

Cloud Computing has been envisioned as the next-generation architecture of IT Enterprise. In the cloud, the data is transferred among the server and client. High speed is the important issue in networking. Cloud security is the current discussion in the IT world. This thesis helps in securing the data without affecting the network layers and protecting the data from unauthorized entries into the server, the data is secured in server based on users' choice of security method so that data is given high secure priority. Cloud Computing has been selected as the next generation architecture of IT Enterprise. In contrast to traditional solutions, where the IT services are under proper physical, logical and personnel controls, Cloud Computing moves the application software and databases to the large data centers, where the management of the data and services may not be fully trustworthy. This unique attribute, however, poses many new security challenges which have not been well understood.

In this thesis, we focus on cloud data storage and transmission security, which has always been an important aspect of quality of service. To ensure the correctness of users' data in the cloud, we propose an effective and flexible distributed scheme with two salient features, Steganography and compression opposing to its predecessors .Cloud storage enables users to remotely store their data and enjoy the on-demand high quality cloud applications without the burden of local hardware and software management. This article explores the barriers and solutions to providing a trustworthy cloud computing environment.

Pooja Lathwal

MCE-635-2K15

A Novel Approach for Test Case Prioritization using Artificial Bee Colony with Genetic Algorithm

Regression testing makes sure that upgradation of software in terms of adding new features or for bug fixing purposes should not hamper previously working functionalities. Whenever a software is upgraded or modified, a set of test cases are run on each of its functions to assure that the change to that function is not affecting other parts of the software that were previously running flawlessly. For achieving this, all existing test cases need to run as well as new test cases might be required to be created. It is not feasible to reexecute every test case for all the functions of a given software, because if there is a large number of test cases to be run, then a lot of time and effort would be required. This problem can be addressed by prioritizing test cases. Software testing is the most important part of software development life cycle. There are various types in software testing which have their own different functionalities. Among them regression testing is most useful functional type of testing which is done in the software maintenance phase. This testing is used to check the errors when any change is made in the existing system. To make system efficient and effective, techniques of test case prioritization are used. The reduction in the cost of testing and fault detection capabilities of testing should be done by test case prioritization.

Regression testing is a critical however to a great degree expensive and time consuming procedure. Due to constrained assets practically speaking, experiment prioritization concentrates on the change of testing effectiveness. Notwithstanding, customary experiment prioritization strategies stress just a single time testing without considering immense information created in Regression testing. This thesis proposes a way to deal with organizing experiments in view of Artificial Bee Colony (ABC) and Genetic Algorithm. The fault detection capacity of an organized test suite is enhanced up to 15% utilizing Genetic Algorithm which shapes the base calculations for prioritization. The adequacy of the calculation is illustrated, and the consequences of tests demonstrate the calculation advances the experiment orderings adequately.

Pradeep

MCE-636-2K15

Evaluation of inferential Accuracy of Open Source Data Set

An important step in machine learning is creating or finding suitable data for training and testing an algorithm. Working with a good data set will help you to avoid or notice errors in your algorithm and improve the results of your application. As creating your own dataset is a very time consuming task in most cases.

If you are interested in practicing applied machine learning, you need datasets on which to practice. This problem can stop you dead. Which dataset should you use? Should you collect your own or use one off the shelf? Which one and why? So Accuracy of Dataset plays a vital in machine learning task like classification, clustering and regression to get accurate and precise result .Also which algorithm to choose to perform the task is also major concern.

The main focus of this thesis to address the above stated problem by checking the quality of UCI dataset. For this purpose 77 Dataset has been taken and five algorithms has been applied with help of Weka tool. Results show which dataset is useful for commercial usage also comparison between the algorithms has been done.

Sapna

MCE-637-2K15

Duplicate Document Detection

World Wide Web (WWW) has a number of web pages. But all the pages on internet don't have the unique information. Some of the documents may have the same content. The content may be exactly same or a part of the document may be similar. Such documents are called as the duplicate or near duplicate documents. If they contains the exact same content, they are duplicate otherwise, if they have some portion of the document similar to other, they comes under the category of near duplicate. Such pages on the web can degrade the relevance of the search engine. Relevance is a measure, how much the retrieved set of document satisfies user need. Relevance can be measured from efficiency and effectiveness of the retrieved documents. Effectiveness represents the ability of the system to retrieve the documents which satisfy user need and suppressing the documents which don't satisfy user need. Whereas Efficiency is the measure how long the system takes to return the results. Thus whenever a document is to be uploaded on the internet, the search engine must check its duplicacy with the existing documents in the index. If the document is a copy of the existing document then it should be discarded from inclusion in the index. However, if the content of new document does not match with the existing document then it must be added to the index. Identification of duplicates or near duplicate documents in a set of documents is one of the major problems in information retrieval. Several methods to detect those documents have been proposed but their relevance is still an issue.

This thesis successfully addresses the significant challenges and has resolved many of them. The main focus of the thesis was to find such a technique for duplicate documents uploaded on web that can efficiently find such documents on the web. In this work, similarity of the newly uploaded document to the already existing documents on the web is calculated and the performance of the technique is measured in form of the speed and rank. The experimental results show that this technique exhibits very good efficiency in handling duplicate documents.

Sneha

MCE-638-2K15

Agriculture Price Forecasting using Time Series Algorithm

Information Mining is rising examination field in Agriculture trim yield investigation. Agribusiness division assumes an essential part of our economy. Farming is the fundamental division contributing in our nation GDP. In the event that our nation's horticulture contributing so much then why ought not we need to concentrate on their improvement? In our nation, the principle issue is that greater parts of ranchers are uneducated and they can't know the informed related strategies they did agribusiness just on the premise of their past experience and their past information. More than 70 rate of the populace relies on upon agribusiness or farming practice. At the point when ranchers gather their yield, they likewise create agrarian crude information. This information can be further utilized as a part of information digging strategies for better harvest administration.

There are different information mining procedures have been created and effectively used in horticulture information mining assignment. Information Mining is developing examination field in Agriculture edit yield investigation.

We consider the issue of foreseeing yield generation. Yield forecast is an essential agrarian issue that remaining parts to be illuminated in the light of the accessible information. The issue of yield expectation can be understood by utilizing Data Mining procedures. This work goes for finding reasonable information models that accomplish a high precision and a high sweeping statement as far as yield expectation capacities.

Ajay Singh Chauhan

MIT-602-2K15

Hashtag Segmentation and Context Dependency using Sentiment Analysis

When purchasing a product for the first time one usually needs to choose among several products with similar characteristics. Companies use to promote their brands and products pointing out good characteristics avoiding to mention the poor ones. The best way to choose the most suitable product is to rely upon the opinions of others.

To extract sentiment about an object from this huge web, automated opinion mining system is thus needed. By devising an accurate method to identify the sentiments behind any text, one can predict the mood of the people regarding a particular product or service.

However, with so much social media available on the web sentiment analysis is now considered as a big data task. Hence the conventional sentiment analysis approaches fails to efficiently handle the vast amount of sentiment data available now a days.

This thesis, successfully addresses the significant challenges and has resolved many of them. The main focus of the thesis was to find such a technique for analysing sentiments posted on web that can efficiently perform sentiment analysis on big data sets and a technique that can categorize the text as positive, negative and neutral in a fast and accurate manner.

In this work, sentiment analysis is performed on a large data set of tweets using eclipse and the performance of the technique is measured in form of the speed and accuracy. The experimental results show that this technique exhibits very good efficiency in handling big sentiment data sets.

Deepa Shekhawat

MIT-604-2K15

Design of Genetic Algorithms based Approach for Opinion Mining with Emojis

With the advancement in technology in almost every field of life, people are so busy in their work that they don't have time for themselves. They depend on the web for buying any product, before that they go through the reviews existing on the web which are very large in amount and sometimes the reviews are present in the form of emoticons. So here comes the technique which extracts the exact opinion about any entity from the web. The technique is called as Opinion Mining. In this field various researches is going on different areas such as feature extraction, classification.

Today, people express their opinions in the different forms and in this paper we are handling some of these different forms and that are-

First is opinions in the form of emoticons, it is a challenging task for an opinion mining system to extract opinions by understanding these emoticons.

Second, one more new feature is added in some social networking sites this feature is using the picture as their comments i.e. we can use pictures with written text on them to describe our opinions. So we can add this in our opinion mining system to put our step forward towards the advancement in opinion mining. This work can be done by using OCR i.e. Optical Character Recognition. OCR is a technique by which we can extract text from the images. OCR have vast field of applications in scanning, PDF file to DOC file convertor etc.

In this paper, we are extracting the opinions based on these emoticons used by users. Also text is extracted from the images, which is a new approach in the context of opinion mining. We are using Genetic Algorithm for extracting the features from the web and neural networks for their classification. Genetic Algorithm helps in finding the optimum features which will lead to optimized opinion mining.

Divya Kaushik

MIT-605-2K15

Context Extraction for Followees Recommendation in Twitter

Twitter is a microblogging social network platform where users share information about their personal lives, engage with other users, follow their favourite news organization, bloggers and much more. Twitter has 300 million active users and approximately 500 million tweets are shared everyday. Due to enormously increasing volume of messages and increasing number of users , information overload has become a serious problem.

This introduces a need for recommender system. With the vast amount of information growing, challenge is to follow the right person in order to get the maximum from twitter. In this study, our aim is to help active Twitter users to find people with similar interests and countervail the information overload problem.

In this work, we have proposed a Context Extraction model for Followees Recommendation in Twitter. The approach is followed by mining user tweets to extract the context representing his/her interest. The Recommendation system starts by crawling user's timeline and collecting the last 100 tweets for further analysis. The tweets are then processed further by structure analyser which generates the triplet consisting of connector word and phrases connected by that connector word. The context extractor analyses the triplets and extracts the salient contexts by accessing the wordnet. Weights are assigned to salient contexts to filter out the irrelevant contexts and top contexts are generated. User clustering is performed for finding similar twitter users and the ranking model recommends the top-k followees to a particular user. The experimental results show that the proposed approach performed effectively.

Divya Sharma

MIT-606-2K15

Implementation of Security measure in Cognitive Radio network against Object Function Attack

Cognitive Radio Technology provides a solution to current spectrum usage inefficiency dependent on its ability to dynamically adapt operating frequencies. Cognitive radio is a technology that can solve the wireless spectrum under-utilization problem by allowing secondary users to opportunistically access the licensed channels without causing interference to the communications of the primary users. Cognitive radio can change its transmitter parameters based on interaction with the environment in which it operates. There are two main characteristics of cognitive radios. The first is cognitive capability, which refers to the ability of the radio technology to sense information from its radio environment. Through this capability, the spectrum resources that are not used by primary users can be detected. Consequently, the best spectrum allocation schemes and transmission parameters can be selected. The second is re-configurability, which enables a user to change the transmitting channel quickly and adaptively according to the radio environment. In a cognitive radio network there are mainly two schemes for secondary users to share the spectrum resources. In one scheme the secondary users reuse the spectrum that is not used by the primary users. The other scheme, called spectrum sharing, allows the secondary users to transmit concurrently with the primary users as long as they do not harm the transmission of the primary users. While cognitive radio is an efficient technique to relieve the pressure of wireless spectrum scarcity, at the same time the characteristics of cognitive radios have introduced entirely new types of security threats and challenges in networks. Since the primary users and the secondary users coexist in the same network, both of them need to be protected, and they are more vulnerable to security attacks compared to the traditional wireless networks without using cognitive radios. Therefore, providing strong security protections is one of the most important requirements for cognitive radio networks. Cognitive Radio Network is an emerging field and have abundant scope of research in it. It can take the concepts like “spectrum sharing”, “wireless body area network”, “Internet of Things” to a next level.

Monika Choudhary

MIT-607-2K15

Design of efficient Routing protocol using power awareness in ADHOC Network

Today, Wireless networks are the most popular and broad field of research work. From the different available wireless networks Mobile ad-hoc network is more popular. Wireless networks have two categories infrastructure based and infrastructure less. MANET comes under the infrastructure less networks hence there is the problem of finding efficient path between source and destination. Mobile ad-hoc networks are the networks that basically consist of mobile node connected to each other through the wireless links. In this each node plays the dual role of acting both as a router and host in order to forward, send and receive packets to each other node in the networks. The mobile nodes are free to move anywhere and organize themselves into the networks when they come in vicinity range of a node. To forward the packets in the mobile ad-hoc networks the routing protocol is needed.

The routing protocols in the Mobile ad-hoc Networks are classified into three categories: Proactive protocol, Reactive protocol and Hybrid protocol. Proactive protocols attempt to maintain consistent, up-to-date routing information between every pair of nodes in the networks by propagating route update at regular intervals. The primary characteristics of this approach are that every node maintains route information to other node in the networks all the time.

On the other hand Reactive routing protocol is also known as On-demand routing technique which takes a very different approach to routing than other classes of protocol. There is no need to maintain a table like proactive protocol and route discovery phase is started whenever required. The benefit of this approach is that signaling overhead is likely to be reduced compared to proactive approaches in particular cases like low moderate low. Third class of routing seeks to be combination of both proactive and reactive approaches.

A wireless network is a network where nodes are free to move in any direction. Mobile Ad-hoc network is a kind of wireless network. Ad-hoc On Demand Distance Vector (AODV) is the most popular reactive routing protocol used in mobile Ad-hoc network for discovering routes. Since wireless network works on dynamic topology, various problems arise in the construction and maintenance of routes among nodes in the network. Another important

aspect of Ad-hoc network is to make efficient utilization of energy resources. If some nodes die early due to lack of energy, they would not be able to participate in the communication as well as create network partitioning. So, power awareness is also an important factor along with bandwidth and efficient route discovery.

To overcome the above said problem, a new protocol has been proposed that will find a path between Source and Destination with minimum number of hops and maximum available power along the path.

Neha Gupta

MIT-608-2K15

Automatic Testing of web Applicaions using Selenium and Jmeter

In the software development applications, one of the major challenges of success or failure is the performance evaluation. Software engineering is not only about the software development processes but also about the effective delivery, deployment and maintenance of the application to the actual users. Software engineering benchmarks like performance, scalability, security, usability and fault tolerance can also be employed in the cloud environment.

The main purpose of this thesis is to automate the web applications on multiple browsers like an end user could do to check its functionality or how a dynamic web application can be automated to check its functionality. The study has focused on a number of the precept factors of performance measures of ecommerce sites and it additionally explored how internet applications are tested for performance and the way it is evaluated. In addition, to identify how well the application performs in relation to the performance objectives. It explores the state of the art in performance and regression testing of web applications.

Functional test scripts are the first scripts created while automating an application in any web based project. Functional testing is the first phase & Performance testing is the last phase in automation project. Functional test scripts can be used for performance testing, when the functional test scripts are created in Selenium.

In particular, use of functional test tools such as Selenium or WebDriver can help the performance test designer to build more intuitive test scripts that lift the level of abstraction from the HTTP request to that of a single user. This approach can greatly reduce the complexity of the test scripts, rendering performance testing a less daunting and less time-consuming process. The benefit of Selenium for functional testing using Junit Framework, provides the ability to utilize the same scripts for Performance testing in JMeter. Which makes the Open Source Integration of Functional to Performance test that is Selenium Integration with JMeter.

Nisha Jha

MIT-609-2K15

Click Prediction using Data Analytics

In software development projects, one of the main challenges of success or failure is how data is handled. It is very important to predict user demand by going through his actions. It can be achieved by machine learning models which are developed using machine learning techniques which are learning algorithms. The target of performing this complex computation on huge dataset is to determine what important features actually contribute to users clicking an ad. In one of the challenges on Kaggle, Outbrain asks to predict which pieces of content its global base of users is likely to click on. I picked their dataset as my project and apply machine learning techniques to find best accuracy contained model.

Further by using classifiers such as Naïve Bayes, KNN, Random Forest, Boosting and SVM helped us to train models based on these features. These classifiers allowed us to show that as the classifier became more advanced with regard to better handling complex data the accuracy increased for this multi-dimensional dataset. This can be very well seen in the results of the various classifiers which showed how the classifiers like Random Forest and Boosting with SVM were able to achieve high accuracy with this data.

This thesis, successfully addresses the significant challenges and has resolved many of them. The main focus of the thesis was to find best accuracy model that can efficiently perform prediction analysis on huge data sets. In this work, sentiment analysis is performed on a huge data set of Outbrain site given by Kaggle using R and the performance of the technique is measured in form of the accuracy using ROC curve.

Ojasvini Dhingra

MIT-610-2K15

Explicit Trust Calculation in Online Social Networks

With the invention of internet, the uses of social networking has increased in almost each and every field. By the use of social networks people can send multimedia objects to other people. There are various popular social networking sites which become part of human's daily routine like Facebook, MySpace, Linkedln, Youtube and various email sites. Communication through social sites is more comfortable and inexpensive. One user can communicate with more than one users at one time through the use of social networking sites without being physically at one location. With the advantages of social networking, it also have some disadvantages like unauthorized access, misuse of personal information, stalking, trolling, privacy theft, spamming etc.

In social networks it may be the case that a lot of the end-users (agents) are usually physically unknown with each other. In this case if two unknown participants wish to communicate with each other for various reasons, the evaluation of their trustworthiness along a certain trust path between them within the social network is mandatory. But the level of trustworthiness may vary and it is sometimes subjective and depends on the person's specific role within the network. It is not an easy task as trust cannot easily be defined through mathematical formulas and algorithmic procedures. Trust may rely on several factors from psychological and sociological factors to computer security factors. There were many attempts to define trust for the online social networks and each covered one or more specific trust factors.

The existing methods for trust calculation resolved many of the problem related to the trust calculation in OSN but also have some issues. They are lacking in using the properties of trust like propagation, trust decay. They are also have problems related to handling large number of users registered with social network. The proposed systemeliminates the problem generated in existing systems. The proposed system finds out all possible trust paths between two users having two main properties: MTL (Maximum Trust Links) and MTT (Minimum Trust Threshold). The generated paths are the shortest path whose length are equal or less than MTL and their trust values are greater or equal than MTT. Then, trust calculation is performed for each trust path by propagating trust from one node to another node considering

possible trust decay at each and every intermediate node. The shortest path with maximum trust values is the result of the proposed method.

The proposed system is successfully implemented and compared with the existing systems. The result analysis shows how the deficiency of existing systems is removed in proposed system. The result shows that proposed system is able to handle the large number of users present in a social network, search shortest paths between source and destination, and decide a possible trust decay at each intermediate node in a path, propagate trust values from previous node to next node and combine more than one trust path to generate a single trust path.

Pooja Rani

MIT-611-2K15

Parallel frequent itemset mining using Interval Intersaction and Bit Vector Approach

As technology is advances, the amount of data used is also increases. This data can be very useful to companies to make business related decisions and to enhance their business by making customer luring policies. But due to its huge size it will be difficult to find useful information manually from these databases. So, Data mining technique is used. It is an inference process which is used to extract information from massive amount of data which is previously unknown and useful for customers. Frequent Itemset Mining is an important data mining method with many real life applications i.e. market basket analysis, catalog design, retail industry, fraud detection, biological data analysis etc. It is a non-supervised process which is used to find patterns or itemsets from large amount of data. It extracts knowledge in the form of repeated patterns. It was first proposed by R. Agrawal in 1993 for market basket. It further proceeds to Association rule mining which is the process of finding associations between items in transactional data. It is user centric as the objective is elicitation of useful rules from which new knowledge can be derived.

In this thesis, classification of frequent itemset mining algorithms is provided with their comparative study and a scalable approach is proposed which provide same result as Apriori but with less scans on dataset and less time and space complexity. In this approach partioning algorithm, bit vector algorithm and incremental mining is used to generate frequent itemsets. Partition algorithm is used for large sized datasets and it saves memory constraint and execution time due to parallel proceeding of partitions. Bit vector algorithm uses vertical data representation and provides fast result by using compression technique. Incremental mining is used to support dynamic dataset. Incremental mining deals with generating association rules based on available knowledge (obtained from mining of previously stored databases) and incremented databases only, without scanning the previously mined databases again. In this proposed algorithm is also compared with Apriori algorithm with an example illustration.

Sadia Parveen

MIT-612-2K15

Sentiment Analysis & Summarization of Reviews in Food Domain

Opinion mining is the study that dissects an individual's opinions, sentiments, perspective, and emotions from written language. It is one of the most dynamic and broadly considered research areas in natural language processing in data mining, Web mining, and text mining. Opinion Mining and Sentiment Analysis is an extension of Data Mining that extracts and examines the unstructured data automatically. It examines and analyses huge amount of unstructured data. It is also known as Sentiment analysis. Opinion mining refers to extraction of those lines or phrase in the raw and huge data which express an opinion or review about something. "Opinions" are key influencers of people's. To extract sentiment about an object from this huge web, opinion mining system i.e. an algorithmic method of analysis of large number of reviews is thus needed. By devising an accurate method to identify the sentiments behind any text, one can predict the mood of the public towards a particular product or service.

Hence, Customer Opinions play a very crucial role in daily life. When we have to take a decision, opinions of other individuals are also considered. Business organizations and corporate organizations are always eager to find consumer or individual views regarding their products, support and services. People are usually interested to seek positive and negative opinions containing likes and dislikes, shared by users for features of particular product or service. Hence, product features or aspects play an important role in sentiment analysis. However, with so much of customer reviews available on the web, it's difficult to read thousands of reviews available online and form an opinion about the product and its services. So there is a need of summarization of these opinions in a way which is less time consuming and helps form an opinion about the product. The main focus of this thesis is to find such a technique for sentiment analysis of food reviews posted on website and categorize them as positive or negative in a fast and accurate manner. Thereafter, forming a summarized view of the opinions using two different techniques.

Shefali Raina

MIT-613-2K15

A Technique for Review Summarization based on Sentiment Analysis

Sentiment Analysis and Opinion Mining is the computational study of user opinion to analyze the social, psychological, philosophical, behavior and perception of an individual person or a group of people about a product, policy, services and specific situations using Machine learning technique. Machine learning for text analysis technically has always been very challenging as its main goal is to make computers able to learn and automatically generate emotions like a human as it is practically very useful in real life scenarios. After a boom in web 2.0 technology this field became the most interesting area of researchers because social media has grown as the fastest medium for availability of opinions. There are many commercial tools available in the market and many researchers have proposed their solutions for opinion extraction, but still there are some problems of text classification and sentiment extraction in opinion mining. These problems arise due to different behaviors, manners and textual habits of users. A sentence can be positive for one, but it may have a negative impact on others so it will be a problem for a machine to generate its emotions. A negative sentence can be written in a positive manner like "What a great camera!", It consumes more battery power, this sentence has a negative opinion about a camera, but it consists only 6 positive keywords. There are mainly four predominating problems viz. subjectivity classification, word sentiment classification, document sentiment classification and opinion extraction. Data mining algorithms are easy to implement, but concludes to poor accuracy meanwhile the machine learning techniques provides better accuracy, but requires a lot of training time, so there should be a hybrid technique which has the advantages of both the techniques. This survey focuses on various Data Mining algorithms, Machine Learning techniques and a brief review stating the comparative analysis of these algorithms. We have followed a systematic literature review process to conduct this survey and also mentioned the future aspects of sentiment analysis and opinion mining.

Due to the widespread and enormous use of the World Wide Web, the amount of data related to a given data is huge. People have become quite liberal in expressing their opinions online. This has led to the existence of thousands of comments, reviews, ratings, etc. on an item/commodity/person and the number is continuously increasing. Humans often rely on the comments and experiences of others before making a purchasing decision. Generally, the first

hand experiences are generally more sought after rather than the seller's description of the product. So if a system could be developed that gives the detailed summary of the various opinions existing online then this can greatly help the potential buyer. Moreover companies selling their products can also use this to track how well the products are being received by the users. So there is a need to develop a mechanism to generate this summary which can enable a naïve user to efficiently use and benefit from this information. However, there are different aspects to consider while developing these summaries. Firstly, it is different from comments while constructing the summary. Moreover in the authors provide us with the insight that the different summaries might contain different levels of information. In such a scenario finding the interesting information is crucial text summarization in which the system simply reproduces a compact version of the original document. Thus a new orientation called Opinion Mining & Summarization has emerged to deal with the problem. Aspect-Based(Feature- Based) Opinion Summarization is one of these summarization techniques which provide brief yet most relevant information about different features related to the target product. Hence the approach is in great demand nowadays because it exactly shows what a customer usually tries to search while referring the reviews. So we focus on extraction of different kinds of features associated with a target entity. Current state of the art suggests that concrete techniques are highly required for identification of those features which are not clearly mentioned. Thus our prime target is to deliver a succinct solution for feature based on the opinion words encountered in user reviews. Finally, summarization of sentences containing both kinds of aspects and generating textual summary of results and finally generate the result with the help of visualization.

Suruchi Garg

MIT-614-2K15

A Novel Architecture for Opinion Mining using Star Ratings

With growing of internet the dependency of humans on the internet has been increased. In every work internet showing its importance either it is professional work or personal needs and choices. Today before purchasing anything or using any services we consult for some of its features. Internet plays very big role in it. It provides many applications serving mankind, and one of them is opinion mining.

In opinion mining where someone is giving their opinions, some are looking for these opinions these are people who want to purchase new product also these opinions are important for the manufacturer of this product who needs to know about their product so that the can evaluate success of that object also they can improve any problem occurs in this product.

In early times when a company needs to know about their product, they do survey which is very time and money consuming task. Opinion mining makes it very simple for all companies to get reviews about their products.

Due to its usefulness today there are many sites which provides interface to interact with users for taking and showing the reviews. Many E-commerce sites have covered the world and are giving best services to the user. These sites provide filtered information that satisfies the user's requirements from the large amount of data and minimize the efforts which users have to apply is opinion mining system is not there. But it also has some limitations and one of them is its working with star-ratings. As many websites take reviews from their user in the form star-ratings. But rating is given for the complete products not for the individual features of the product. Which make it hard to mine exact opinion with respect to user's requirement?

In this thesis I am going to explain my work in this area, here I am making a system which takes opinion of users in form of star-ratings and on the basis of those data this system will tell either that product is according to user's need or not.

Tamanna Sachdeva

MIT-615-2K15

A Novel Approach for Sentiment Analysis Using Deep Learning

In recent years, the development of internet and web 2.0 technologies, enabled by cost reduction of technological infrastructure, has been an exponential increase in the amount of information in online systems. These very large volumes of information are very difficult to process by individuals, leading to information overload and affecting decision-making processes in organizations. Therefore, providing new techniques for creation of knowledge is important in organizational strategy. This large amount of information on web platforms make them viable for use as data sources, in applications based on opinion mining and sentiment analysis.

Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. Opinions are central to almost all human activities and are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are, to a considerable degree, conditioned upon how others see and evaluate the world. For this reason, when we need to make a decision we often seek out the opinions of others. This is not only true for individuals but also true for organizations. The opinions of others have a significant influence in our daily decision-making process. These decisions range from buying a product such as a smart phone to making investments to choosing a school—all decisions that affect various aspects of our daily life. Before the Internet, people would seek opinions on products and services from sources such as friends, relatives, or consumer reports. However, in the Internet era, it is much easier to collect diverse opinions from different people around the world. People look to review sites, e-commerce sites, online opinion sites and social media to get feedback on how a particular product or service may be perceived in the market.

Vrinda

(MIT-617-2K15)

Travel based Recommendation System based on Collaborative Filtering

Due to exponential growth of internet, everyone is dependent on internet to get even a small suggestion. It has become a big concern to provide the better suggestions to users. As more and more services turning online, recommendation is also a form of online suggestion where a better result of recommendation can be fetch in minimum time. To provide better suggestions in less time, travel based recommendation system plays important role.

Data has been collected from twitter. Check-in of user timeline has been extracted to get the visited locations of the users. Locations has been ranked according to number of users visited these locations. Then, location dependence calculates the co-existence between users and form the final ranking of the locations. It provides the better filtering of data that removes the cold start problem and provides only a set necessary data.

To show these sequence of ranked locations markers has been form in a sequence of coexistence of locations and provides suggestion in the form of complete set of locations. In this travel based recommendation system a sequence of locations have been provided as a result of suggestions on Google map with the help of fetched data from twitter timeline.

Yashasvee Shukla

MIT-618-2K15

Smart document clustering (To remove document duplicacy) using TF-IDF & cosine similarity Approach

In recent days document clustering becomes important in information retrieval system, text mining, data mining, web analysis and many other application. Document clustering is an automatic clustering operation of text documents in which related documents are shown in the same cluster and different or unrelated documents are shown in different cluster. Now a day document clustering becomes a significant technique for speedy and efficient information retrieval on web. Document clustering is the efficient way to find the nearest neighbors of a document**.**

In this work, smart document clustering is performed using tf-idf and cosine similarity to remove the duplicacy of the document. By using this approach we can prevent a duplicate document to upload on the website. In this approach the document which we want to upload on the website will first compare to whole of the documents which are existing on the website if the document match more than the cosine similarity limit then the document will not be upload else the document will upload. We can see the comparison result after press the compare button. We are facing an ever increasing volume of text documents. The abundant texts flowing over the Internet, huge collections of documents in digital libraries and repositories, and digitized personal information such as blog articles and emails are piling up quickly every day. These have brought challenges for the effective and efficient organization of text documents. Clustering in general is an important and useful technique that automatically organizes a collection with a substantial number of data objects into a much smaller number of coherent groups. In the particular scenario of text documents, clustering has proven to be an effective approach for quite some time—and an interesting research problem as well. It is becoming even more interesting and demanding with the development of the World Wide Web and the evolution of Web 2.0.

In this work, we generate and implement checker's algorithm which deals with the duplicacy of the document content with the rest of the documents in the website.

Arif Khan

MNW-641-2k15

Content Extraction from Social Media

Internet is a source of huge information. By just one click on your laptop or mobile phone you can gather information about anything in the world. E-commerce, social networking, education, entertainment, online tourism etc are widely used web applications. These application providers allow users to share their feedbacks, reviews and suggestions online on their websites. Social media platforms such as Twitter, Facebook, and blogs have emerged as valuable - in fact, the de-facto virtual town halls for people to discover, report, share and communicate with others about various types of events. These events range from widely-known events such as the U.S Presidential debate to smaller scale, local events such as a local Halloween block party. During these events, it is often witness a large amount of commentary contributed by crowds on social media. This burst of social media responses surges with the “second-screen" behavior and greatly enriches the user experience when interacting with the event and people's awareness of an event. Monitoring and analyzing this rich and continuous flow of user-generated content can yield unprecedentedly valuable information about the event, since these responses usually offer far more rich and powerful views about the event that mainstream news simply could not achieve. Despite these benefits, social media also tends to be noisy, chaotic, and overwhelming, posing challenges to users in seeking and distilling high quality content from that noise. Event Extraction is a Natural language processing or Information Extraction task that involves identifying instances of specific types of events and their associated attributes.

Extraction of events with its specific attributes is most common challenge faced during event extraction. This thesis presents event extraction system using supervised learning technique. The system is able to classifying the events attributes using Naive Bayesian machine learning technique and give the result in tabular form. In the process of constructing the system, many previous works are reviewed and some of them are applied to the thesis and different methods are compared for reaching a best solution.

Anjali Pandey

MNW-642-2k15

A Novel Approach for PCA based Summarization

Opinions play a pivotal role in decision making in the society. Opinions and suggestions given by others are the base for an individual or a company while making decision. Summarization of opinions is a prime and important step in Opinion mining. With the explosion of the abundant data present on social media, it has become important to analyze the text for seeking information. There is huge amount of data on the web which contains the redundant information. Summarization of dispensable content thus is a necessity. Summarization has to be performed in such a manner such that the text is condensed in the shortest way possible and important information is extracted from the text by preserving its properties without any lossless condensing. With the huge amount of reviews posted online, a summary is thus required to influence a person in making correct decision considering all its important thematic words into account.The task of automatic summarization has increased its interest among communities of NLP and Text mining. Reviews summarization using PCA is incorporated in our methodology as it finds relevant thematic words in the dataset and by combining the essential parts of the review containing relevant thematic word, Summaries thus generated are extraction based summaries and are well formed and structured to convey gist of the text.

However, with huge amount of content available on the web opinion mining is now considered as a big data task. The main focus of the thesis was to find thematic words from the dataset based on open mind common-sense knowledge taken out from ConceptNet assertions, analyze the sentiments used for that thematic words then assigning the scores from SentiWordNet and forming the matrix ,applying PCA and generating the summary that consist of all essential reviews.

Deeksha

MNW-643-2K15

Secure cloud based data dissemination in VANETs

Vehicular Ad hoc Networks (VANET) is a subclass of Mobile Ad Hoc Networks (MANETs). It is the most advanced technology that provides Intelligent Transportation System (ITS) in wireless communication among vehicles and between vehicles and Road Side Equipment (RSUs) according to IEEE 802.11p standard. VANET provides broad range of safety and non-safety applications. In this all vehicles participating in the network turn themselves into a wireless node. These nodes can detect each other with the help of sensors which are inbuilt in vehicles. As the vehicles pose dynamic nature, VANET is likely to face stale entries and congestion. Due to high mobility of vehicles, topology of the network changes frequently. Hence, this will reduce the network life time and increases routing overhead.

In order to avoid these kind of problems, many solutions have been proposed of which clustering is one of the most effective way of managing and stabilizing such networks. Clustering in VANETs is the process of organizing vehicles into groups based on some rules, criteria or common characteristics. Clustering reduces the messages count flow and increases the connectivity in the network. However, there are some issues related to cluster formation**.** Extensive literature survey results no such proposal which uses bidirectional traffic model. In this model all vehicles, that are assumed to be approaching the RSUs from both directions, are considered. Data Dissemination, Routing, QoS parameters, and Security techniques with respect to vehicle's mobility are open research issues in this type of networks.

To overcome the problems faced in the clustering based approaches for secure communication in VANETs, a scheme is proposed which makes use of cloud for data filtering and its dissemination among interested vehicles. The main objective of the proposed scheme is to decrease congestion on roads and maintain a smooth traffic flow. The proposed scheme will be helpful to prevent traffic jams which further increases road capacity. This scheme allows vehicular nodes to access traffic related information from the cloud of that region.

Deepti

MNW-644-2K15

Domain Specific Feature Extraction for efficient Opinion Mining

In olden days people were only information consumers but since the arrival of Web 2.0 they plays more important role in publishing information on Web in the form of comments and reviews. The user generated content forced the organization to pay attention towards analyzing this content for better visualization of public's opinion. Opinion mining or Sentiment analysis is an autonomous text analysis and summarization system for reviews available on Web. Opinion mining aims for distinguishing the emotions expressed within the reviews, classifying them into positive or negative and summarizing into the form that is quickly understood by users. Feature based opinion mining performs fine-grain analysis by recognizing individual features of an object upon which user has expressed opinion.

In the context of Opinion Mining from customer reviews, machine-learning approaches have been recommended; however, it is still a very challenging task. In this thesis, work has addressed the problem of Opinion Mining, and proposed a Natural Language Processing approach that undertakes Dependency Parsing, Pre-processing, Lemmatization, and part of speech tagging of natural texts in order to obtain the structure of sentences by means of a dependency relation rule. This extends traditional dependency parsing to phrase level. This concept is then implemented for extracting relations between product features and expressions of opinions This extraction approach is evaluated using customer product reviews collected from Amazon for nine different products. Based on analysis, it is found that the proposed dependency patterns provided a moderate increase in accurate results than the baseline models. This study also found that the average percent change for aspect and opinion extraction was significantly improved compared to the baseline models. The results are shown for study and discussed how they relate to comparative experimental results. This work ends with a discussion that highlights the strong and weak points of this method, as well as direction for future work. Examples are provided to demonstrate the effectiveness of using Dependency Relations for optimizing the problem of Opinion Mining.

Divya

MNW-646-2K15

Hybrid Approach for Efficient Opinion Mining

The advent of Web 2.0 has led to an increase in the amount of sentimental content available in the Web. Such content is often found in social media web sites in the form of movie or product reviews, user comments, testimonials, messages in discussion forums etc. Timely discovery of the sentimental or opinionated web content has a number of advantages, the most important of all being monetization.

Understanding of the sentiments of human masses towards different entities and products enables better services for contextual advertisements, recommendation systems and analysis of market trends.

The focus of this research is a combined approach for opinion mining. This thesis presents the combination of SVM and KNN classification algorithm and find out how much given text sound good. The proposed technique is compared with other existing techniques specially Naïve Bayes' and results shows that the proposed technique is better as compared to the other technique.

Jyoti

MNW-647-2K15

Secure Data Aggregation in Wireless Sensor Networks

A wireless sensor network (WSN) generates huge amount of data with limited resource constraints which are susceptible to communication failure and various types of security attacks due to broadcast nature of communication. Data aggregation techniques are used to gather data from different sources to eliminate redundant and duplicate data. There are various issues in data aggregation such as delay, redundancy elimination, accuracy, traffic load and security. Security is a major concern in data aggregation to achieve data confidentiality, aggregator node availability, data integrity, and sensor node authentication. A sensor node which is providing data to the aggregator node must be an authentic node. Aggregator node must be capable of detecting false data from unauthorized nodes in order to avoid flooding of unnecessary data which may lead to denial of service attack.

The proposed scheme for secure data aggregation technique is using digital watermarking for achieving data integrity. The scheme also carries a solution for vulnerability by providing an initial trust estimate based on authentication of the sensor node before allowing it to participate for communication and transmission in the area governed by a particular base station.

Digital watermarking is a mark embedded in data which is used for tracing ownership of the signal. It helps to verify integrity and authenticity of sensed data. These types of watermarking concepts are used by sensing nodes as well as aggregator node to protect data integrity. In watermarking, each sensor node embeds a unique watermark to sensed data so that base station can verify for data integrity and combating data modification, data deletion and false data insertion attack. The objective is to achieve data integrity and data authentication of the sensed data by use of watermarking and public key infrastructure by providing a secure data aggregation without significant loss of data.

Masoom

MNW-648-2K15

Improved Focused Crawler using Graphical Property, Web Page Classifier and Link Evaluation

Internet is a source of huge information. By just one click on your laptop or mobile phone you can gather information about anything in the world. E-commerce, social networking, education, entertainment, online tourism etc are widely used web applications. When the user write the Query on the web browser to find out the specific topic or query in the true database. The database are very big and contains the large amount of information about everything in the world. In database contains the Meta data also. To find the specified information in that database contains large amount of time to find that relevant information i.e. requirement of crawler is occur.

Web crawler are the program that collect the information from the internet .Web crawlers are a central part of search engines, and details on their algorithms and architecture. Multi thread in web crawler is faster in collection data than a single thread web crawler. A focused crawler is topic-specific and aims selectively to collect web pages that are relevant to a given topic from the Internet. However, the performance of the current focused crawling can easily suffer the impact of the environments of web pages and multiple topic web pages.

The thesis is mainly developed to provide best quality of improved focused crawler. A Focused crawler finds the keywords and place name to the given topic or specified topic is interested in find the high relevance of the web page. This information is used to improve the services to makes the efficient web crawler. The geographically property of the place name are used to find the relevant place name quickly compared to the tyonym ontology. This improved focused web crawler are also helps to identify the high relevance web pages or low relevance web page using the web page classifier. At the end result is shown in form of output in figures using the various technique to describe positive and negative result related this. The improved focused crawler can use the high relevance of that web page.

Preeti

MNW-649-2K15

Review Spam Detection Combined Approach of Behavioral Analysis and Structural Analysis

Online reviews play a significant role in today's ecommerce. Most of the customers now a days are depending on the reviews and ratings for taking decisions of what to buy and from where to buy. Thus, Pervasive spam, fake and malicious reviews are affecting the decisions of customers while buying products. These reviews also affects stores rating and impression. Without proper protection, spam reviews will cause gradual loss of credibility of the reviews and corrupt the entire online review systems eventually. Therefore, review spam detection is considered as the first step towards securing the online review systems. Main aim to give overview of existing detection approaches in a systematic way, define key research issues, and articulate future research challenges and opportunities for review spam detection. Opinion spam (or fake review) detection has attracted significant research attention in recent years. The problem is far from solved. In this survey, here presents various methods of opinion spam detection. Merchants selling products on the Web often ask their customers to review the products that they have purchased and the associated services. As e-commerce is becoming more and more popular, the number of customer reviews that a product receives grows rapidly. For a popular product, the number of reviews can

be in hundreds or even thousands. This makes it difficult for a potential customer to read them to make an informed decision on whether to purchase the product. It also makes it difficult for the manufacturer of the product to keep track and to manage customer opinions.

This thesis, contains combined approach of behavioral analysis and structural approach for increasing result. To detect the review spam is necessary to for exact result. Because for a customer accuracy matters more, show there should be genuine reviews to satisfy them. Here tried to increase the accuracy for expected results

Reena

MNW-650-2K15

A Study and Development of Gesture Control Mechanism for Handicapped using MATLAB

In Engineering and language technology Gesture recognition is a phenomenon with the goal of interpreting human gestures via mathematical algorithms. Gestures can originate from any bodily motion or state but commonly originate from the face or hand. Here we are focusing on the hand gesture recognition with the help of Sixth Sense Technology. With the help of gesture recognition humans can communicate with the machine (HMI) and interact naturally without any mechanical devices. It is possible to point a finger at the computer screen so that the cursor will move accordingly using the concept of gesture recognition. The advantage of gesture control is that the user not only can communicate from a distance, but need have no physical contact with the computer.

The Sixth Sense is a wearable gestural interface that augments the physical world with digital information and lets us use natural hand gestures to interact with that information. The color marker tracking technique of Sixth Sense technology is been used so as to make the interaction possible with the handicapped people including handless people for whom other gesture recognition techniques cannot fulfil their requirement as they need hand postures and positions to fire responding functions. With color tracking technique of Sixth Sense technology the benefit of gesture recognition is possible for handless handicapped people also.

Shivam Sharma

MNW-652-2K15

A Novel Approach for Novelty Detection via Topic Modeling

Novelty detection is the technique which is used to fetch the sentences/document having new information than the previous sentences/document Novelty detection is the technique used to extract novel information from a set of relevant documents or from a same document in a given topic (query). Our approach works on the concept of topic modeling. A topic model is a type of statistical model for discovering the abstract “topics ” that occur in a collection of documents. Topic modeling is a frequently used text- mining tool for discovery of hidden semantics structures in a text body. Topic models can help to organize the large collection of unstructured text bodies. Here, Topic refers to a set of words that frequently occur together.

And after having a survey on Novelty detection, In this thesis we implemented an approach to generate the novel document from the given set of documents. The idea is a combination of two techniques k-means clustering and Sshlda (semi supervised hierarchical latent dirichlet allocation). The presented idea provides better results than the existed ideas.

Shrishti Vashist

(MNW-653-2K15)

Ontology based Extraction using Web Mining

Internet is a source of huge information. By just one click on your laptop or mobile phone you can gather information about anything in the world. A system that processes a set of web pages and extracts information regarding education, job, exams, country, politics, Bollywood or any kind of information can be given is an example of information extraction system. Information extraction (IE) is that which retrieve limited and relevant information from natural language data and as well as unstructured data. Sometime it is processing a human language text which is also known as natural language processing (NLP). Information extraction is that which required any type of information. Ontologies are domain specific; it means that different domains are having different Ontologies, in which it shows the relationship between the different classes and entities. Ontologies are application dependent and it can create specific ontology for specific application. Ontologies -hierarchical representation which shows the classes and sub-classes relationship of components of any domain specific concept which helps to extract the information on the basis of related concepts and it also building and updating Ontologies. Ontology based information extraction (OBIE) is a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract that types of information and gives the output by using ontologies. OBIE is a form of knowledge extraction where the knowledge is based on ontology. Ontology construction is generally not connected with information extraction; it can be seen as an important step in the ontology-based information extraction process. The system is able to classifying the rating and server comparison using Naive Bayesian machine learning technique and give positive or negative rating about a go daddy, blue host and big rock. In this, we use the Naive Bayesian for the output.

In this thesis, it extracts the relevant information from structured and unstructured data. And it extracts the positive and negative rating of the data. This rating of data is used to make it easy. In this thesis, it also works on server comparisons like pricing, operating system and type of the system.

Shweta

MNW-654-2K15

Semantic Similarity between two Documents using topic MAP

Computing semantic similarity between any two entities (word, sentences, and documents) is a crucial tasks on the web .Semantic Similarity plays a significant and big role in information retrieval(IR), natural language Processing(NLP) and many other tasks of IR related tasks such as relation extraction, and document clustering,. It is a concept where a pair of documents is measured to computing the Semantic Similarity between documents using various similarity measures. Computing similarity between a pair of documents with efficient method is really a major difficult task for the user. Similarity measure those are used to find similarity, assign a real number between 0 and 1 to a pair of documents. If both documents are similar then user will get a numerical value 1 otherwise they will get 0.This Thesis proposes a framework for computing the semantic similarity between documents based on topic**.** We have also added a NLP parser .In this Thesis work, first we have done the documents pre-processing process. TF-IDF weighting schemes are used.

The process starts with pre-processing of the documents using NLP parser. Then Topic map is build that represent the document in compact form and cosine similarity measures is used to measure the similarity between these topic maps.

Sonam

MNW-656-2K15

Design of MD5 based Authentication Protocol using Mobile Agent

Wireless networks has been experiencing an explosive growth similar to the Internet, this is largely due to the attractive flexibility of anytime, anywhere network access enjoyed by both users and service provider. While the emergence new wireless technologies can enable truly ubiquitous Internet access, it also raises issues regarding the dependability of the Internet service delivered to users, which may be impacted by the time-varying channel, limited spectrum, mobility, and particularly the security .Basically Wireless Local Area Network (WLAN) can operate in two modes, the infrastructure based and the Adhoc networks. Many organizations are deploying the infrastructure based wireless network to provide connectivity to locations with difficult terrain, poor accessibility, or places difficult to reach by direct cabling, to complement the existing wired networks. A lot of attention has been given to the provision of these wireless networks, but little attention has been given to the provision of adequate security for the emerging wireless networks making the networks prone to traditional link-layer attacks readily available due to proximity. Wireless network security is more concentrated and complex than security of wired network because wireless is broadcast in nature, making it possible for anyone within the range of a wireless device to eavesdrop and intercept the packets sent without interrupting the flow of data between the wireless device and the access point over the air. User authentication is a reliable wireless security protocol for best safeguard against the risk of unauthorized access to the wireless network.

Mobile Agents (MA) is an effective paradigm for distributed applications and is particularly attractive in a dynamic network environment involving partially connected computing elements. MA is defined as a software component which is either a thread or a code carrying its execution state to perform the network function or an application. MA can act as a middleware and perform network and other application related functions based on underlying infrastructure: fixed wire network, wireless cellular network or mobile ad hoc network. MA paradigm is an emerging technology for developing applications in open, distributed and heterogeneous environment like the Internet.

Sudarshan Thakur

MNW-657-2K15

Relevant content Extraction & Text Summarization

The amount of textual and multimedia information on world wide web has been increasing many folds every year. A user seeking information on the web is often overloaded with colossal amount of related documents by search engines and information retrieval systems to satisfy his information need. In this context, it has become increasingly important to develop information access systems that provide focused and precise answers to the user. Text Summarization is a popular information access solution for information overload problem.

Text Summarization is a process of extracting or collecting important information from original text and providing that information in the form of summary. It is an important research area in today's era of the fast growing information age. As information is growing day by day on the internet, it is difficult for users to identify the relevant information. Users have to read the whole document to determine whether the given document is relevant or not. With the help of text summarization, a shorter version of large text documents can be produced by keeping the relevant information from the original text document as well. It basically condenses the salient features from the text by preserving the content and serves the meaningful summary. Classification can be done in two ways: extractive and abstractive summarization. Extractive summarization uses statistical and linguistic features to determine the important features and fuse them into a shorter version. Whereas abstractive summarization understands the whole document and then generates the summary. In this thesis, focus is mainly on the extractive summarization techniques, as they are easy to compute and quite successful till now.

As, we focus only on extractive methods of summarization, where only the text units from document collection are used in producing a summary. Sentence is considered as a basic text unit for the summary. Extractive summarizers generally follow a sequential framework, that include Pre-processing of text for sentence boundary identification and extraction, a Feature Extraction stage where several statistical, linguistic and heuristic models are employed to estimate sentence importance, Sentence Ranking stage that estimates sentence importance through weighted linear combination of the features and finally Summary Extraction, during which a subset of ranked sentences are selected into summary.

And after having a brief survey on extractive text summarization techniques, a novel technique for text summarization or we can say a technique for generating the summary of a given document using Naïve Bayes algorithm has been proposed. The experimental results show that this technique exhibits good efficiency in handling big data sets thus generating the summarized text.

We can say that in this thesis, we are going to study a variant of text summarization, that is our "Proposed Summarization", focusing only on relevant and novel information, and produce an informative, non-redundant summary of the topic.

Yashaswi

MNW-658-2K15