CONTENTS

Automat. Comput. Appl. Math.
Volume 22 (2013), Number 2

COMPUTER SCIENCE

ALIN IOAN TURCU, GABRIEL DRAGOMIR
A Modern Approach to Intelligent Transportation Systems Development. 207

ION GIOSAN, EMŐKE OLTI, SERGIU NEDEVSCHI
Traffic Road Obstacles Detection Based on Analysis of Relative Motion Vectors. 217

MIHAELA DÎNSOREANU
Supervised Approaches to Sentiment Analysis 225

MIHAELA DÎNSOREANU
Unsupervised Methods for Sentiment Classification. A Case Study for Twitter. 235

ADRIAN PECULEA, BOGDAN IANCU, VASILE TEODOR DĂDÂRLAT, EMIL CEBUC, ALEXANDRU COSTAN
Framework for Testing Proxy Mobile IPv6 Functionalities. 243

CIPRIAN POCOL, SERGIU NEDEVSCHI, ION GIOSAN
Obstacle Detection Based on Single Frame Stereo Vision. 247

CALIN CENAN
A Proposed Architecture to Support Web Based Ontology Extraction. 257

A Modern Approach to Intelligent Transportation Systems development

Turcu Alin Ioan, Gabriel Dragomir
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
alin.ioan.turcu@gmail.com, gabriel.dragomir@cs.utcluj.ro

Abstract—**The application domain of this paper is Intelligent Transportation Systems (ITS). These systems can be used to provide means to analyze traffic, reroute traffic, dynamically adjust the traffic light and speeds limits. This domain suffers from a shortage of data sources that could provide an accurate description of the traffic situation. Typically, to provide routing and navigation solutions map attributes in the form of static weights are derived from road categories and speed limits and used for calculating the best route. With the arrival of Floating Car Data (FCD) and specifically the GPSbased tracking data component, we now have the means to derive accurate and up to date traffic data with leads to a more qualitative traffic information. This work presents a system that facilitates the collection of FCD based on mobile devices, produces dynamic travel time information, and provides traffic prediction services based on the traffic information. In order to increase the relevance of the predictions the proposed system analyzes besides traffic data also the time component (time of day, day of week, if there is a holiday or not. Etc) and the weather conditions. The proposed system uses a partitioned datawarehouse that facilitates the use of paralleled algorithms.**

Keywords-GPS, FCD, DW

I. INTRODUCTION

In general, an ITS relies on location-based information: It monitors and processes the location of a certain number of vehicles (used as probes) to obtain information on estimated travel time, driving conditions, and traffic incidents. Using a relatively large amount of probes, the early stages of bottlenecks can be detected, and traffic can be directed to other routes to mitigate congestion and provide more expedient and efficient itineraries to travelers.

A variety of sensors can be used to obtain traffic information. These traditionally fall into two main categories: fixed sensors and Global Positioning System (GPS) receivers. Fixed sensors, which can be either hard-wired or wireless, are used outdoors on roads and can provide information on the speed at which vehicles are traveling, as well as the capacity of a given road. Fixed sensors that are inductive, piezoelectric, or magnetic can be placed under the road, and those that use radar, ultrasonic, or infrared with pyroelectric effect can be placed beside it. Each sensor is equipped with a control unit, battery system, solar panel, and transmission system.

GPS receivers are used inside vehicles and periodically transmit information on a vehicle's location and speed, together with the associated GPS time, to a central system. These receivers are composed of a control unit and a transmission system, which is based on text messages and/or the transmission of data packets.

Both fixed sensors and GPS receivers contain limitations. The installation of wired fixed sensors is costly, as they require a cable network to transmit information to the central collecting point. Moreover, they provide only point-based information on traffic conditions, and to gain a realistic and complete view of automobile traffic in a certain area, a large quantity of sensors must be installed. In this scenario floating car data (FCD), becomes an important tool for traffic assessment and prediction [1]. Floating car data (FCD) refers to using data generated by one vehicle as a sample to assess to overall traffic conditions. Having large amounts of vehicles collecting such data for a given spatial area such as a city (e.g. taxis, public transport, utility vehicles, and private vehicles) will create an accurate picture of the traffic condition in time and space.

While GPS receivers provide better data, they also pose significant problems. If each street is to have sufficient sensing capability, a significant percentage of vehicles must be equipped with GPS devices and mass deployment is expensive.

In the past few years, research activities have begun to explore the usage of data obtained from mobile cellular networks. Mobile phone positioning techniques generally provide less accuracy than the GPS, but the wide diffusion of mobile equipment (ME) in addition to the widespread installation GSM transmitters, in both urban and rural areas makes such positioning techniques very appealing. Mobile devices location data are abundant and if they are aggregated and processed, they can effectively be used to understand driving behaviors, identify areas of congestion, and many others. Mobile phones are, in short, a wide-area sensor network, the data from which could complement and integrate with those coming from traditional sensor networks.

Several projects regarding feasibility and field tests of ME location-based ITSs have been developed in Europe (e.g. France [2] and U.K. [3]) and North America (e.g. California [4], [5] and Washington, DC).

This work presents the technology needed to collect FCD from vehicle fleets, to derive road-network-related travel times (map-matching), to provide efficient data

Traffic road obstacles detection based on analysis of relative motion vectors

Ion Giosan, Emőke Olti, Sergiu Nedevschi
Computer Science Department
Technical University
Cluj-Napoca, Romania
ion.giosan@cs.utcluj.ro, emokeolti91@gmail.com, sergiu.nedevschi@cs.utcluj.ro

***Abstract*—Computer vision obstacle detection on either road lanes or sidewalks is very important for traffic participants. This paper presents an approach for obstacle detection from sequences of consecutive monocular color image frames. Key-points are uniformly distributed in a grid structure on each input image. A Lucas-Kanade optical flow algorithm is performed between each pair of consecutive frames, on the considered key-points, in order to compute the relative motion vectors. Background movement estimation is computed across frames using a RANSAC procedure. Optical flow vectors that are belonging to the background are filtered out. The others are considered to be within the obstacles and are grouped by a hierarchical clustering algorithm in separate obstacles by analyzing their locations, angles and magnitudes. Spurious clusters with low number of motion vectors are filtered out. Finally, an imminent collision warning is issued both visually and acoustically when an obstacle is detected to be too close and it is about to crash with the ego-camera.**

Keywords-obstacle detection; motion vectors; hierarchical clustering; background movement estimation; collision alert.

I. INTRODUCTION

Nowadays, the number of intelligent vehicles and smart devices is growing rapidly due to the technological possibilities that are into a continuous development process. In case of an intelligent vehicle, the driver is alerted by a driving assistance system when there appear potentially dangerous situations. Usually it includes many safety functions like obstacle collision warning, lane departure warning, lane keeping assistance, speed keeping assistance, etc. There are also other traffic participants like vision deficiency persons [1] who can be guided or warned on the sidewalks by a similar system. This could be possible if they carry on a smart device with a specific application installed which performs environment understanding.

Here resides the motivation for building high accuracy obstacle detection module that can help either drivers or persons with vision deficiency or even blind people. An obstacle detection module must determine the regions of interest from traffic scene where exist obstacles. It may also provide this information to a subsequent module which may classify it into a specific obstacle class like pedestrians, poles, trees, vehicles, wall etc. Another important aspect for the detected obstacles is to find their exact location within the traffic scene or at least to infer if there may be an imminent collision with the ego-vehicle or with the person who carries the smart-device.

There are many different technologies like LASER-scanners, RADAR, infrared sensors, ultrasound sensors and video cameras that can be used for obtaining the scene information. However, video cameras are capable of acquiring visual information that can be further processed for environment understanding. This way of getting the traffic scene information is no pollutant for the environment, being also similar to the human eyes vision system. Depth computation of scene elements is also very important. This can be done basically by using a stereo-cameras setup. We don’t use a stereo-setup due to the fact that it implies higher costs and usually smart-devices don’t have integrated stereo cameras. We use a single color camera setup, with low costs, which offers us sufficient information for implementing all the processing operations.

We propose an approach for detecting the obstacles either on lanes or sidewalks by analyzing the relative motion vectors between frames. Optical flow is computed on uniformly distributed grid points by Lucas-Kanade algorithm. The result defines the relative motion vectors which are then used for both background movement estimation and obstacles segmentation. A RANSAC algorithm is used for determining the background movement in order to be filtered out from the motion vectors field. The obstacles are defined by clusters of motion vectors obtained after a two-step hierarchical clustering procedure considering their specific features: location, magnitude, and angle. Noisy clusters are finally removed obtaining the valid obstacles. The possibility of imminent collision of each obstacle with the ego-camera is evaluated and signalized. We also present the evaluation of the entire obstacle detection system.

II. RELATED WORK

Obstacle detection from a video sequence is one of the most important task used in many real time automotive applications. The researchers carry out a lot of work for developing very good solutions for obstacle detection from both monocular and stereo vision images.

In case of stereo vision systems the task is easier than in monocular systems. Stereo systems acquire much more traffic scene information using at least two cameras. This allows the obstacle detection [2] to be done by analyzing both

Supervised approaches for sentiment analysis

Mihaela Dinsoreanu
Computer Science Department
Technical Unversity of Cluj-Napoca
Cluj-Napoca, Romania
Mihaela.dinsoreanu@cs.utcluj.ro

***Abstract*—The sentiment analysis problem is highly relevant nowadays since the availability of huge amounts of user generated content on virtually any domain. The task is not a trivial one involving several challenges and also possible approaches, none of which proved so far to be general enough to be applied in any domain, language or on any data. In this paper we present an extensive state-of-the art review on the challenges and supervised solutions in the field and our results so far in the problem of opinion extraction, both from English and Romanian documents.**

Keywords-sentiment analysis; opinion extraction; supervised machine learning; feature selection

I. Introduction

Sentiment analysis is one of the hottest topics in the information retrieval field since the availability of huge amounts of user generated content. Users are now able to express themselves on any kind of topic (products/services, politics, etc.) in various types of media such as articles, blogs, comments, reviews, tweets, Facebook postings etc. The resulted content can reveal useful information not only for individual users that are searching for a certain product or service but also for companies to understand the global opinion of their customers related to their products or to tune their campaigns according to the user feedback. For example, if a new product receives several bad reviews on a certain aspect, the company might react timely to improve that aspect before the sales of that product drop significantly.

The main issue of the sentiment analysis field is to correctly identify the semantic of the user generated content in terms of opinions out of natural language text that might be grammatically incorrect or might contain special symbols like hashtags, emoticons etc. Moreover, not any user generated content might express opinions or sentiments so the first challenge is to discriminate between opinion bearing and non-opinion bearing content. A first discrimination would make sense between objective and subjective content. Objective content represents facts that normally do not involve any sentiment. Even in case of subjective content, it might not involve sentiments or opinions, just user perceptions. Therefore, identifying correctly user expressed opinions is not a trivial problem. A first step towards opinion identification is to understand what we are looking for. In other words we should know what the components of an opinion are in order to recognize them in texts. A widely adopted structure for opinions was proposed in [1] and [2] as a quintuple ***<e,a,s,h,t>*** where ***e*** represents the addressed entity/target, ***a*** represents a particular aspect of the entity ***e*** the opinion is about, ***s*** is the actual sentiment on the aspect ***a***, ***h*** is the holder of the opinion and ***t*** is the time the opinion was expressed.

Obviously, these opinion components are seldom explicit and clear in the text. In many cases we have to deal with implicit components such as implicit aspects. For example in the sentence “The printer is not expensive but the ink cartridge is” the involved aspect of the printer is the price but it is not explicitly mentioned. Other implicit opinions are related to pronouns that represent previously mentioned entities that can be targets or holders, or comparative opinions. In this case the opinion does not provide an absolute evaluation of the target/aspect but a relative one to another product.

In our work so far we addressed some facets of the sentiment analysis problems and obtained relevant results.

The rest of the paper is structured as follows: the next section provides a brief review of the main facets of the sentiment analysis problem. It is followed by a review of the most relevant approaches found in literature. Section IV presents a survey of our proposed solutions highlighting the most significant results while the last section concludes the paper.

II. Problem facets

Opinion mining is a challenging task that involves various types of problems, for each problem several approaches have been reported in literature. We are briefly mentioning some of them in the next sections.

A. *Problem facets*

The main task of sentiment analysis is the accurate identification of opinions with all the components defined above from different types of raw input text. This task is not trivial since it involves not only the identification of the sentiment polarity but also associating it to the correct entity or aspect.

One facet of the problem is the *granularity* level. Research results found in literature are addressing three granularity levels: document, sentence and aspect level. We present in next section existing approaches for each of these levels. Another facet of the problem is the *language* issue. Most of the approaches are addressing documents written in English. Several high quality tools and data resources have been developed for the English language so far. There are some approaches that handle other languages such as Chinese, Japanese, Arabic etc. but there is no proven solution to be language independent or even to perform similar in several languages. Moreover, in the framework of the same language, even English, we can distinguish between correctly written documents (e.g. articles, news) and documents that

Unsupervised methods for sentiment classification. A case study for Twitter

Mihaela Dinsoreanu
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
mihaela.dinsoreanu@cs.utcluj.ro

***Abstract*—The sentiment analysis problem is highly relevant nowadays since the availability of huge amounts of user generated content on virtually any domain. The task is not a trivial one involving several challenges and also possible approaches, none of which proved so far to be general enough to be applied in any domain, language or on any data. In this paper we start with an extensive state-of-the art review on the unsupervised solutions in the field as well as the available lexical, semantic and benchmark data resources. We present our unsupervised sentiment classification approach applied on tweets, and our results applied on four benchmark datasets. Since our results are comparable with other supervised approaches we conclude that our solution is relevant.**

Keywords-sentiment analysis; opinion extraction; unsupervised machine learning; Twitter

I. Introduction

Sentiment analysis is not necessarily a very new topic but is increasingly relevant since the availability of very large amounts of user generated content. The most common way to inform ourselves about the quality of a certain product/service is to analyze the user reviews on that product/service. Companies are able to access and to analyze valuable user feedback that can drive their future strategies much quicker than in the past. However, reviews are not the only means for people to express themselves, people are now eager to share their experiences and feelings with their friends in real time on social networks such as Facebook or Twitter, too. Although there are many research efforts in the field of sentiment analysis, not all the approaches are suitable for social network content. The reasons for that are manifold: (i) one aspect is the fact that the objective of a review is to provide opinions on a target while in social network postings, such as tweets, people might just want to share news/events, to express their emotions like happiness or sadness that are not necessarily oriented towards a certain target (ii) reviews are usually written correctly, in complete sentences that allow for traditional Natural Language Processing (NLP). On the other hand, tweets have a limited size (140 characters) and usually contain slang, abbreviations, special symbols, emoticons etc.

Most of the sentiment analysis approaches are aiming to identify the basic components of an opinion: the target (i.e. the identity that is referred), the opinion holder (who expresses the opinion) and the opinion itself. In the case of social network content or postings the holder is by default the current user but, although the posting involves a sentiment polarity, there is not always a target. The user might feel happy because she had a great time or feels loved etc. In terms of sentiment analysis methods, most of the approaches employ a supervised method that involves the existence of annotated datasets that are used to train classifiers in order to correctly classify other new datasets. These approaches usually yield a performant classification especially if applied in related domains. However, supervised approaches are not that useful to classify tweets since there is no relevant annotated corpus and it is very expensive in terms of effort to create such corpus. Moreover, given the diversity of content existing in tweets, supervised approaches are not able to perform acceptable to classify them.

In our work so far we investigated unsupervised approaches for polarity classification of tweets. We also analyzed existing lexical, semantic and benchmark data resources that can help to increase the classification performance. We proposed a processing flow and applied our solution to four benchmark datasets in order to compare our results with other state-of-the art solutions.

The rest of the paper is structured as follows: the next section provides an overview of the related work addressing unsupervised sentiment classification. A comparative discussion on the existing resources follows next. Section IV presents the main steps of our proposed solution highlighting the most significant results while the last section concludes the paper.

II. Unsupervised approaches for sentiment analysis

A. *Sentiment analysis*

The next section is concerned with an overview of the unsupervised approaches for sentiment analysis applied on three levels of granularity: document, sentence and aspect.

1) Document level approaches

The first significant unsupervised approach for sentiment classification was presented by Turney in [1]. The approach starts with a POS-tagging operation targeting the identification of adjectives and adverbs. Based on a set of 5 combinations of POS elements (e.g. adjective + noun) the method identifies bi- or tri-grams that are likely to express sentiments. The extracted phrases get an estimated sentiment orientation based on the PMI-IR algorithm that uses mutual information as a measure of the strength of semantic association between two words. Turney uses as reference

Framework for testing Proxy Mobile IPv6 functionalities

Adrian Peculea, Bogdan Iancu, Vasile Teodor Dădârlat, Emil Cebuc, Alexandru Costan
Computer Science Department
Technical University of Cluj-Napoca, 400114, România
Email: {Adrian.Peculea, Bogdan.Iancu, Vasile.Dadarlat, Emil.Cebuc}@cs.utcluj.ro

***Abstract*—Mobile IP networks are able to provide mobility for telecommuters, without application-level interruptions. Proxy Mobile IPv6 is used for network-based IP Mobility management to Mobile Nodes. The proposed framework for testing Proxy Mobile IPv6 functionalities is implemented in GNS3 - that uses emulators to run various operating systems as in real networks and connects to real networks through the network interface cards. In this way the accuracy of the framework is close to the one of real networks. The framework topology allows for a general approach of the Proxy Mobile IPv6 feature testing. The experimental results proved that the Mobile Node roaming in the Proxy Mobile IPv6 domain maintains its IPv6 addresses and the communication with other nodes is not interrupted.**

***Keywords*—Mobile IPv6, proxy, mobility, framework, testing**

I. Introduction

Internet Protocol version 6 (IPv6) was developed by Internet Engineering Task Force (IETF) to solve the exhaustion of Internet Protocol version 4 (IPv4) addresses problem, due to massive Internet penetration of fixed and mobile nodes.

IPv6 has brought many changes on the new Internet protocol as described in [1]. Expanded Addressing Capabilities was achieved by using a 128-bit address, adding a new type of address (anycast address) and improving multicast addressing. The header has a simplified format, many fields became optional or were dropped, to reduce packet overhead. Also IPv6 added enhanced support for extension headers and options.

As mentioned in [2] IPv6 added several enhancement in respect to IPv4, other than a larger address space and a new header format. Autoconfiguration is an important feature of IPv6, allowing hosts to receive an IP address with or without the presence of a DHCP server, in a plug-and-play manner. IPv6 also provides support for real-time services, security support and enhanced routing functionality. Of a great importance is the support for nodes mobility in IPv6.

With the ever-increasing Internet penetration and telecommuters working from virtual offices, mobile IP (MIP) can provide continuous connectivity for users applications, regardless of their location. Using MIP host devices are able have a permanent IP address, thus allowing seamlessly support for application level connections without interrupting the connections.

A mobile IP network consists of the mobile node (MN), the home network (HN) - where the mobile node receives the IP address, a home agent (HA) responsible for forwarding the packets and where the MN registers its location, and the care-of address (CA) where the MN registers its attachment to a foreign agent (FA) in the new network.

Mobile IPv6 (MIPv6) represents the mobility support protocol for IPv6 and it is a natural evolution from mobile IPv4 (MIPv4). MIPv6 is defined in RFC6275 [3] as a protocol that allows nodes to remain reachable while moving around in the IPv6 Internet and allows a mobile node to communicate with all IPv6 nodes, mobile or stationary. Deploying PMIPv6 requires that mobility management processes should be done by the network side on behalf of the users devices. The paper proposes the use of a framework for testing Mobile IPv6 functionalities, in order to assure a proper deployment in a real world scenario.

The paper is organized as follows: Section II provides background information and theoretical considerations about Mobile IPv6 components and operations. Section III presents the proposed framework for testing Proxy Mobile IPv6 functionalities, implemented in GNS3. Section IV presents the experimental results and performance measurements, used to test the proposed framework. Section V concludes the paper and discusses future research topics, in the context of Future Internet and mobile environments.

II. Theoretical considerations

Proxy Mobile IPv6 (PMIPv6) [4] provides network-based IP Mobility management to a Mobile Node (MN). While Mobile IPv6 (MIPv6) requires the participation of the MN in IP mobility-related signaling, PMIPv6 entities track the movements of the MN, initiate the mobility signaling, and set up the required routing state without requiring the participation of the MN in any IP mobility-related signaling. The functional entities of PMIPv6 are presented in figure 1.

Mobile Access Gateway (MAG) performs mobility-related signaling on behalf of the MN. It obtains an IP address from Local Mobility Anchor (LMA) and assigns it to MN, retains the IP address of the MN when the MN roams across MAGs

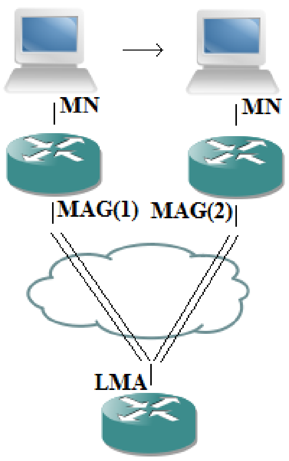

Fig. 1. PMIPv6 functional entities

Obstacle Detection Based on Single Frame Stereo Vision

Ciprian Pocol, Sergiu Nedevschi, Ion Giosan
Department of Computer Science
Technical University of Cluj-Napoca, Romania
{Ciprian.Pocol, Sergiu.Nedevschi, Ion.Giosan}@cs.utcluj.ro

Abstract—**The obstacle detection from single stereo frames is a less investigated topic, while it is more tempting to add temporal information, like optical-flow (low-level) and obstacle tracking (high-level). A good understanding of obstacle detection in single frames is required for better results in obstacle detection from sequential frames. This survey uses a taxonomy that classifies the approaches based on their main processing space of the depth data. The methods for ground-obstacle separation are briefly detailed as well. At the end, there is a comparative analysis of the processing spaces and of the approaches of different research teams.**

Keywords: obstacle detection, single frame, stereo vision, survey

I. Introduction

One possibility to do 3D measurements in a generic scene, by using a moving camera, is a technique named structure from motion. In the same time, it computes the camera position relative to its position in the previous frame, this way simulating the principle of the stereo vision.

By using two distinct cameras, rigidly mounted on a rig, a real stereo vision system is obtained and it has several advantages over a mono vision system:

- the relative position of the two cameras is always the same and it can be computed with high accuracy by using specific calibration methods;
- the structure of the scene can be determined even when there is no motion;
- the calibration dramatically reduces the search space of features from one image to the other one, increasing the correspondence certainty and the depth precision.

Having the depth information available, it is tempting to go one step further and add optical flow information [1], even though the computation complexity increases significantly. The main objective of this paper is to provide a good understanding of the possibilities to detect obstacles from single stereo frames.

Some approaches use assumptions on the scene structure. For instance, it is assumed that the ground surface is planar, so that the difference of the IPM images (Inverse Perspective Mapping) emphasizes the obstacles [2]. The same assumption is used in [3] in order to quickly reconstruct the ground surface, by reducing the range of possible disparities; the approach is named "ground plane stereo".

This survey mainly focuses on approaches that aim to detect generic obstacles in generic ground scenes (using single stereo frames, as said before).

The obstacle detection and the ground detection may have common or similar parts. That's why the ground detection approaches will be briefly presented as well, because they perform the ground-obstacle separation. Some approaches detect the whole visible ground surface [11], while others detect only the limit of visible free space, which is actually the frontier between the ground surface and the beginning of the obstacles [16].

The source space of the 3D data is the perspective image enriched with depth represented by disparities. It is named disparity map and is also known as the U-V-disparity space; U and V being coordinates on the image plane. The U and V coordinates are defined relative to the optical center of the image [6]. Their resolution can differ from the image resolution: for instance one unit on the U axis = 2 pixels, in order to compress the data. The U-disparity and the V-disparity histograms are often used; they accumulate the pixels having the same (U, disparity) values and (V, disparity) values respectively.

From the disparity map, 3D points can be obtained; they are often represented in a polar or in a Cartesian space. Different approaches often use spaces that are derived from the disparity map or from the 3D space.

Due to the perspective geometry of the image formation, the coordinate system of the disparity map has a polar nature on both the lateral (U axis) and vertical (V axis) directions, while the depth's nature is based on disparities. On the other hand, the Cartesian coordinate system has its X and Y axes parallel with the image's U and V axes, while the Z axis represents the depth (some approaches may use different names for these axes).

Usually, any detection algorithm uses a main space and other secondary spaces:

- in order to use the real perception possibilities offered by the disparity space while using reasoning in the Cartesian space of the scene or vice-versa;
- in order to take into account coordinates that were lost when particular spaces were generated.

II. Approaches

The disparity based depth was firstly used, being the output of the stereo matching process. Later, 3D Cartesian points – metrically expressed – were obtained from the disparity map by the 3D reconstruction step.

In the following, the existing detection approaches are presented. They are classified by the used depth

A proposed architecture to support Web based ontology extraction

Călin CENAN
Dept. of Computer Science
Technical University of Cluj-Napoca
Cluj-Napoca, România
Calin.Cenan@cs.utcluj.ro

Abstract—**Automatic ontology building strategies have been researched in literature, but have not converged to a domain-independent or generally accepted methodology. Our approach propose an architecture and uses a set of finely tuned annotation tools in order to build and populate an ontology, based on natural language processing techniques. We built a system that supports fuzzy entity matching, generated relations based on verb phrases, parenthetical processing and language dependencies. A taxonomical-building module produces class placement and subclass assertions based on gazetteer annotations. Our solution will be applied to automotive forums and we hope that we will obtain a close to completeness ontology with good quality metrics.**

Keywords: ontology building, automotive forums, natural language processing

I. Introduction

In the information era, we are faced with large and even huge amounts of data. It is impossible for us to handle this information properly, as time is not enough for proof-reading these huge collections of data. In many domains, this problem has aggravated so much that experts delegate documentation to other employees, e.g lawyers, who delegate to one or more paralegals. Other domains, like history or medicine have also expanded beyond any prior limit known to man, and require people to read extensively in order to make simple decisions, state simple facts or build accurate statistics. And the amount of data is still growing and becoming more and more complex, unmanageable by experts or, even more so, by laymen.

Much research has gone into finding easy and comfortable ways of structuring and indexing information. Classical methods include artificial orderings, such as the year of emerging of a document or book, author name, publication codes etc., while the newest approach of the so-called Semantic Web aims to order documents by their semantics alone. It is a very difficult task and has not yet become mainstream on Internet, but it shows promising results. At the heart of this web of things there are ontologies, a form of knowledge representation which combines concepts, relations and rules pertaining to a single domain. Building qualitative ontologies is therefore crucial for the quality of information and requires domain expertise. Today, most ontologies consist of manually inputted data and structures, and in the meantime the merging and automatic building of ontologies is rising as a prominent subject of research.

Usually, ontology building is domain-dependent. However, considering the case of automatic ontology building from unstructured text, it is clear that language paradigms are the same cross-domain and should be considered universal. It is possible to devise a methodology for extracting information from unstructured text, regardless of the domain, and add domain-specific processing afterwards. The following paper presents such an endeavor, where language constructs are considered for assertions into ontology. The most common domains that we have observed in the literature were scientific or technical, where the language is semi-structured and the vocabulary is somewhat restricted, but here we also address the challenge of managing noise and subjectivity stripping from a non-technical text. In this paper beside a general domain approach we focus our process on ontology which can be obtained from automotive forums.

The system itself should consist of ontological assertions, OWL ontology rule definition and reasoning, concept fuzzy matching and natural language processing at the level of the technical text found on automotive forums. In these forums the messages and information are organized by categories and subcategories, with threads of discussions (collections, multitudes of messages) which associated a title, a description, a start date and possible the date of the last posting. There are forums of holders of vehicles generally organized by manufacturers and countries (www.daciaclub.ro, vwclub.ro, www.hyundaiownersclub.co.uk, www. hyundaiclub.ro). In general such forums of discussions are organized for a manufacturer (Dacia, VolksWagen, Hyundai, according to the examples above), organized internally by different communities from different countries and also organized by different models available. For the assistance of users in the purchase of a car or the resolution of different problems which may appear in operation, the automatic processing of this data would be useful to evaluate the degree of satisfaction of the owners and the identification of the main technical problems (the most common, which appear in large number). This degree of satisfaction and the identification of the main technical problems are of major interest not only for the users who want to buy a new or second-hand car, but especially for the manufacturers of vehicles who can thus obtain a reliable feedback from their customers.

If we want a processing of this data it is obvious that the elements of these chat forums which are similar to social