Information Extraction Techniques in Different Application Domains

Manuja George
Post-Graduate Student
Department of Computer Science and Engineering
Karunya University, India

J. A. M Rexie
Assistant Professor
Department of Computer Science and Engineering
Karunya University, India

Abstract

Information extraction is the process of extracting relevant information from documents in the internet. Information must be most relevant to the query which is given by the user. Information extraction can be useful in various domains like classifying customer complaint scenario, question answering, pattern induction, company news evaluation or search engines. Now public is increasingly depended on the internet they want the system to think like humans and take appropriate decisions. In order to improve relevant data extraction different technologies are introduced day by day. To manage these kinds of technologies, new skills are also needed. Natural language processing has greatest role in this information extraction. To understand more user oriented information and communication system, natural language processing methods are used. In this survey paper such techniques are discussed. Different techniques are used in different application domains however they all perform information extraction.

IJERT

1. Introduction

Many of the actions we make during the day turn into data for organizations to use for our own profit and learning. Using an automatic teller machine, filling out a form for a driver's license, order a book on the Internet, booking a flight on an airline - all develop into digitized data to be sorted, managed, and used by others. In each of these cases, someone at some time has determined how the data from these users will be received, stored, processed, and made available to others. Organizations and individuals use databases to bring independent sources of data together and store them by electronic means. Thus, a database is gathered of related files that are combined, ordered and stored together. One collection of related files might be appropriate to employee information. Another group of related files might include sports statistics. Organizations and individuals may have and use many diverse databases, depending on the nature of the work consisted. To access and control data in a database Data Base Management Systems (DBMS) are used. A database management system is software pack up that enables users to edit, link and update files as need dictate.

Over the years, many large organizations have accumulated huge amounts of data about their suppliers, customers, products, and services. Even many new Web-based companies have total large databases about people and products as they have developed. The WWW is itself a large distributed data repository with innumerable potential. With the growing awareness that these vast data resources can be tapped for significant commercial gain, curiosity in data mining, data warehousing, and data marts has virtually exploded. Data mining, also known as Knowledge Discovery in Databases (KDD), refers to computer assisted tools and techniques for sifting all the way through and analyzing these infinite data stores in order to find tendencies, patterns, and correlations that can guide decision making and increase understanding. Data mining covers a wide range of uses, from analyzing customer purchases to discovering galaxies. In essence, data mining is the parallel of finding gold nuggets in a peak of data. The monumental task of finding concealed gold depends heavily upon the influence of computers. The purpose of DM is to analyze and realize past tendencies and forecast future tendencies. By predicting future tendencies, business organizations can better position their products and services for monetary gain. Non-profit organizations have also achieved major benefit from data mining, such as in the area of scientific progress.

Nowadays, there have been more researches running for natural language processing. Natural language processing (NLP) is the capability of a computer program to understand human speech as it is spoken. NLP is a part of artificial intelligence (AI). The development of NLP applications is challenging because computers habitually require humans to speak to them in a programming language that is exact, unambiguous and highly structured or, possibly through a limited number of clearly enunciated voice commands. Human speech, however, is not always exact; it is often ambiguous and the linguistic structure can depend on many difficult variables, including slang, regional dialects and social situation. Current approaches to NLP are based on machine learning, a type of artificial intelligence that analyzes and uses patterns in data to progress a program's own understanding. Most of the research being completed on natural language processing revolves about search, particularly enterprise search. Common NLP tasks in software programs today consist of: Sentence fragmentation, part-of-speech tagging and parsing.

Natural language processing can be used for different kind of applications like information extraction, question answering, classifying customer complaint, relation extraction etc. Different kind of approaches used in this applications; bag-of-words approach, shallow parsing and key word based approaches.

2. KEY CONCEPTS

2.1. Data Mining

Extraction of necessary information from mass volume of data is called data mining. Knowledge is exposed from data in which necessary data is taken out from the raw material is called knowledge discovery from data.

2.2. Information Extraction

Information extraction process is process of extracting relevant information or patterns from internet. Information extraction is applied in search engines, classifying customer complaint scenario, question answering, and pattern induction.

3. Analysis of different information extraction techniques

This section contains a study on some of the information extraction techniques in different application domains.

3.1. General framework approach for subjective evaluation

General framework is [1] one information extraction strategy used to handle subjective information. Subjective information extraction is useful in real life applications. It enables system to make human like decisions. Approach is tested in the company news evaluation. The general frame work has four steps: parts of speech tagging, syntactic parsing, generation of relation and criteria evaluation.

The public have more demands and special skills are needed to operate new technologies. To communicate with different stake holders, new methodologies are also needed. Now a days natural language processing techniques are used for user oriented systems. In natural language processing, first the query is syntactically parsed then converted to semantic structure. That will be translated into sql queries. There is one challenge in mapping between semantic structure and data base objects. Two phrases are needed for this purpose. First is query reduction and next is query expansion. In query reduction phase, useless words are dismissed and in query expansion phase the scope of search is expanded. In general frame work approach, information extraction is automatic. It extracts information and gives meaningful interpretation. Most of the information extraction systems do not perform subjective evaluation task. But general framework does subjective evaluation such as grading essays or appraising the merit of company news.

Humans while reading a document read word by word and then understand the meaning and parse the sentences. Then they find out the syntactic relationship. That relationship becomes a unit of information. Unit of information is easier to handle than full text. General frame work approach also has the same steps. First step performing parts of speech tagging. In the second step give syntactic structure to the sentence. In the third step find the relevant syntactic relations and in the fourth step perform criteria evaluation. General frame work approach is useful in company news evaluation. Whenever new news break out, it analyze and present that information to the user.

In the first step, that is parts of speech tagging, assigning of the parts of speech to each word in a sentence is done. Different parts of speech can be used for a word. There is a need to find out the correct parts of speech tag for the word which is more suitable to the context. Second step is syntactic parsing. Syntactic parsing means for each sentence creating a syntactic parse tree. Syntactic parse tree helps to find the grammatical relations between phrases or words. Each sentence can be represented in different syntactic parse tree form. Statistical syntactic parsers need to choose the appropriate tree. Third step is relevant relation extraction. This will be more structured. Relation might be subject-verb, subject-verb-object, adjective-nouns

IJERT

etc. Finally clusters are made. First column contain words, in second column parts of speech and third column cluster number. Last step is criteria evaluation. If number is high, then that word has more positive effect of that relation.

3.2. TEXTRUNNER system for information extraction

Open information extraction [5] is a new extraction, paradigm that extracts sets of large relational tuples. It does not require any human input. Its input is a corpus and output is a large set of extracted relations. KNOWITALL is a web extraction system which uses small sets of domain-independent extraction patterns and labels its own training examples. It uses parts of speech tagger not parser. But it requires large number of quires and downloaded web pages. Open information extraction introduced TEXTRUNNER which is scalable. It is a fully implemented system can extract relational tuples from text.

Input to TEXTRUNNER is corpus and a set of extraction is the output. That set is indexed to support exploration. TEXTRUNNER has mainly three modules: self supervised learner, single pass extractor and redundancy based assessor. Self supervised learner operates in two steps. In first step it labels positive or negative to its own training data. Extractor module uses a Naïve Bayes classifier, which is trained by that labeled data. Parser can train an extractor. So learner uses a parser which can identify trustworthy extractions and label them. In Naïve Bayes classifier, they are used as positive training examples. Extractions are in the form of tuple t=(e_i,ri,$_j$$e_j$,) e_i and e_j are strings. That does not entities. Relationship between them is also a string represented as $r_{i,j}$. First step is to parse the sentences and make dependency graph. Then find all base noun phrases. For each pair of noun phrases, it finds the relation $r_{i,j}$ in the tuple. If certain constraints are met, then learner labels it as positive.

In single pass extractor it tag each word with parts of speech. Then finds out relations. It examines text between noun phrases, and non essential phrases are eliminated. It then finds chunker for each noun phrase which helps to find the word which is part of entity. Finally tuples are applied to classifier will label as trustworthy and which will be stored by TEXTRUNNER.

In redundancy based assessor process, merge the tuples by TEXTRUNNER. Here entities and relations will be identical and the count of sentences from which the extractions are obtained When a user gives a query to access a subset of tuples, by using TEXTRUNNER relevant subsets will be given and irrelevant subsets are hidden from the user. This can be done in seconds.

3.3. Semantic approach to pattern extraction

A novel algorithm [4] is presented for the acquisition of information extraction patterns. It is hypothesized that, useful patterns will have similar meaning. Information extraction is useful in language processing tasks such as identification of lexical patterns and question answering. New algorithm uses WordNet. That is used to generalize the extraction patterns and describe the implementation. Semantic approach used two techniques. They are based on recognition of relevant documents and identification of relevant sentence.

The set of documents can be trained by a given IE scenario which is against the system. Documents are annotated and may be either relevant or irrelevant. Corpus is pre-processed and generated a set of patterns. Patterns represent the sentences in corpus. That set is called as S. Learning process aims to find out the subset that is relevant to IE scenario. User provides relevant patterns called seed and represented as S_{seed}. From this patterns make the accepted patterns S_{acc}, S minus S_{acc} called as candidates represented as S_{cand}.

Score will be assigned to pattern in S_{cand} by a function f that is based on those which are currently in S_{acc}. Real number is assigned to the patterns. Highest score patterns are called S_{learn}, they are suitable or inclusion in accepted pattern. S_{learn} patterns will e added to accept patterns and they are removed from candidate patterns is obtained, process will be stopped. Otherwise repeats from scoring steps.

The function to score the pattern can be done either based on documents that contain relevant patterns which are already identified or based on how their meaning are similar to relevant which is already known.

3.4. COGEX: Logic prover for question answering

COGEX introduced an idea of reasoning that can be automatically being applied to question answering. Question and answer passages are transformed to logic representations by new approach. To understand the relationship between question and answer text, world knowledge axioms and linguistic axioms are supplied to prover [3]. COGEX verifies and extracts all logical relationships between questions and answers by codifying the question answer text and resources.

IJERT

Logic representation of text provides the syntax based relationship which is captured by COGEX. QA logic prover uses world knowledge axioms that provide a link from question concept to answer concepts. WordNet glosses represented in logic forms provide the axioms. COGEX is able to re-rank answer with intelligent representation effectively and efficiently. Answers will be correct and exact.

Input to the COGEX is question in logic representation answer paragraphs, and lexical information. Some other inputs consists of NLP axioms and answer in logic form (ALF). First input goes to axioms builder which converts all the inputs to axioms. Justification of answers starts once the axioms are completed and loaded. Relaxation module is called if proof failed. Relaxation module has two purposes. At the time of parsing of text and transformation of logic form there might be error occurring, that need to be compensating this is one of the purpose. Sometimes NLP axioms can't provide correct answers. Relaxation module detects the correct answers in that case. Relaxation module, reduce the proof score and re-attempt the justification. Until the proof succeeds or proof score is below a threshold value, the justification relaxation loop is continued. Ranked answer and answer justification are the output from COGEX. Output is provided once all candidate answers are processed and according to proof score give answers are ranked.

3.5. Syntactic and semantic generalization approach for sentence classification

Inferring semantic property approach [2] extends the mechanism of logical generalization toward syntactic parse trees and trying to detect semantic signals unobservable in the level of keywords. Generalization from a syntactic parse tree defined by the set of maximum common sub-trees which is obtained and is performed at the level of paragraphs, sentences, individual words and phrases. Approach analyzes the semantic features of this similarity measure and compares it with the semantics of traditional anti-unification of terms. Nearest Neighbor machine learning is then performed to relate the sentence to a semantic class. Inferring approach explore the possibility of high-level semantic classification of natural language sentences based on syntactic parse trees. Approach is inspired by the notion of anti-unification which is able to generalize arbitrary formulas in a formal language. Syntactic generalization is considered to be structure-based and deterministic. It has linguistic features retain their structures and not represented as values. Graph-based machine learning can predict the possibility of complaint scenarios based on their argumentation structures is taken as an assumption in approach.

Inferring semantic properties can be applied in search application to improve their relevancy. In this similarity is measured by computing the syntactic generalization between queries and sentences. Search is performed in two steps. First step is by obtaining documents in the keyword based approach. Then do parse tree generalization techniques and filter the documents which got. Then obtain the results of the conventional search and for the query, each sentence, and each search-hit snapshot calculate the score of the generalization results. The search results are then re-sorted, and return the results to the user those that are syntactically close to the search query are assumed to be relevant.

Inferring semantic properties approach basically has three steps. First it finds out the syntactic parse tree for each sentence and does syntactic generalization. Second step is to find out semantic generalization for the sentences. Third step is to map the syntactic generalization with semantic generalization and find optimal path. Optimal path is the logic form which has highest score.

In syntactic generalization it has more steps. First step is for each sentence find out the parse tree. In parse tree node will represent the word that has lemma or parts of speech. And each node will be connected each other. For each phrase type, split the sentence into sub trees. According to phrase type sub trees are grouped. Equivalence transformation also extended. Each pair of sub tree is generalized. For each calculate the score. Then select the generalization has highest score. In semantic generalization first will give semantic role like agent, patient, instrument or other adjuncts to the words in sentence. In syntactic generalization, semantic role labeling can be served as additional constraints. In information retrieval systems srl helps to reduce the number of passages or documents retrieved.

In mapping step, generalization framework is being combined with semantic representations, like logic forms, to get the text's meaning. Here build the most accurate semantic representation by combining the preset semantic information with learned information. Taking one assumption that the target feature caused by common linguistic features of a training set. In this step make these features on both syntactic levels by means of generalization and on the semantic level by means of logical anti-unification. For this purpose, first proceed on the syntactic level and then check its performance

IJERT

on the semantic level of logic forms. Finally, map the syntactic level generalizations to the semantic level generalizations. The order of generalization is selected by finding the expression which has highest score.

3. Conclusion

Constituency parse tree provides slightly rich set of features compared to a bag-of-words approach or shallow parsing. Need to deal with such a rich set of features with its inherent configuration by a structured machine learning approach. The semantic level of categorization classes is much higher than the level of semantic role labeling or semantic entailment. SLR does not aim to construct complete formal meanings. In this categorization classes, such as suitable phrase extraction and relevant/irrelevant search results, are of a high semantic level. Web-based metrics that calculate the semantic similarity between words or terms are corresponding to the measure of similarity.

4. References

[1] Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni "Open information extraction from the web", in: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence AAAI Press, Hyderabad, India, 2007, 2670 – 2676.
[2] Boris A. Galitsky, Josep Lluis de la Rosa, Gabor Dobor Dobrocsi "Inferring the semantic properties of sentences by mining syntactic parse tree", Data & Knowledge Engineering, 2012, 21-45.
[3] D. Moldovan, C. Clark, S. Harabagiu, and S. Maiorano "Cogex: a logic prover for question answering", in: Proc. of HLTNAACL, 2003.
[4] H. Mangassarian, and Hassan Artail "A general framework for subjective information extraction from unstructured English text", Data & Knowledge Engineering 62, August 2007, 352 –367.
[5] M. Stevenson, and M.A. Greenwood "A semantic approach to IE pattern induction", in: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics ACL, 2005.

IJERT