Efficient Autoscap Algorithm -Automatic Construction of Dictionary

Priyadharshini. M
SREC-Coimbatore,

Saranya.V
SREC-Coimbatore,

Swathi. S
SREC-Coimbatore,

Ramakrishnan. S
SREC-Coimbatore,

Abstract

*Many data mining techniques have been proposed for mining useful patterns in text documents. However how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problem of **polysemy** and **synonymy**. Over the years, people have often held the hypothesis that pattern or phrase-basedapproaches, they all suffer from the problem of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern or phrase-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. This paper presents an innovative and effective pattern discovery technique which includes the processes of clustering techniques and information extraction, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.*

Introduction

Knowledge-based NLP systems depend on a domain-specific dictionary that must be carefully constructed for each domain. Building this dictionary is typically a time-consuming and tedious process that requires many person-hours of effort by highly-skilled people who have extensive experience with the system. Dictionary construction is therefore a major knowledge engineering bottleneck that needs to be addressed in order for information extraction systems to be portable and practical for realworld applications. We have developed a program called AutoScap that automatically constructs a domain-specific dictionary for information extraction. Given a training corpus, AutoScap proposes a set of dictionary entries that are capable of extracting the desired information from the neighbour texts documents[2]. If the training corpuses representative of the targeted texts, the dictionary created by AutoScap will achieve strong performance for information extraction from text documents.

A Review of Related Work

The term- based methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, term- based methods suffer from the problems of **polysemy** and **synonymy**, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want. To overcome this problem **phrase-based approaches, or pattern mining-based approaches** have been proposed, which adopted the concept of closed sequential patterns[9], andpruned no closed patterns. These pattern mining based approaches have shown certain extent improvements on the effectiveness. However, the paradox is that people think pattern-based approaches could be significant alternative, but consequently less significant improvement are made for the effectiveness compared with term based methods.

are hidden huge set of data and interpret then to useful knowledge and Information[5].

The two fundamental issues regarding the effectiveness of pattern-based approaches: **low frequency andMisinterpretation**. Given a specified topic, a highly frequent pattern is usually a general pattern, or a specific pattern of low frequency[9]. If we decrease the minimum support, a lot of noisy patterns would be discovered.

Misinterpretation means the measures used in pattern mining (e.g., “support” and “confidence”) turn out to be not suitable in using discovered patterns to answer what users want. The difficult problem hence is how to use discovered patterns to accurately evaluate the weights of useful features in text documents.

Related Work

1. Knowledge Data Dictionary

Knowledge discovery is a process that extracts implicit potentially useful or previously unknown information from data. It describes[5]: Data comes from variety of sources integrated into a single data store called target data. The data then preprocessed and transformed into a standard format. The data mining algorithm[1] process the data to the output in the form of patterns and rules, then those patterns and rules are interpreted to new or useful knowledge or information**.**

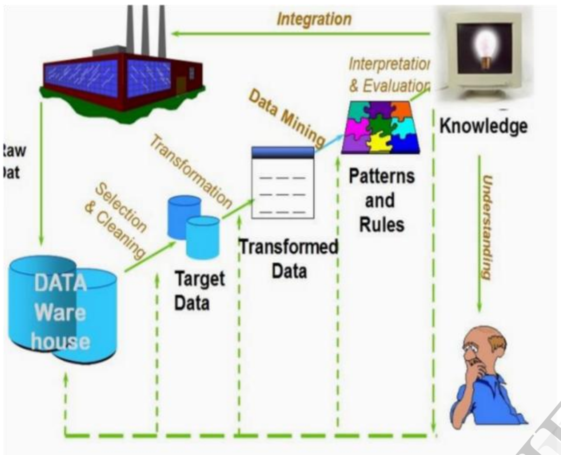

Fig.1.1 Process Of KDD

Goals:The Ultimate goal of knowledge discovery and data mining is to find patternthat are hidden huge set of data and interpret then to useful knowledge and Information[5].

2. Natural Language Processing

Natural language processing (NLP) is a modern computational technology that can help people to understand the meaning of text documents. For a long time, NLP was Struggling for dealing with uncertainties in human languages. Recently, a new conceptbased model was presented to bridge the gap between NLP and text mining, which analysed terms on the sentence and document levels. This model included three components. The first component analysed the semantic structure of sentences; the second component constructed a conceptual ontological graph (COG) to describe the sematic structures; and the last component extracted top concepts based on the first two components to build feature vectors using the standard vector space model. The advantage of the concept-based model is that it can effectively discriminate between no important terms and meaningful terms which describes a sentence meaning. Compare with the above methods, the concept based model usually relies upon its employed NLP techniques.

3. K-Means Cluster

The K-means algorithm takes a input parameter K and partitions a set of n objects into a K-clusters so that the resulting intra-clusters similarity is high but the inter-cluster similarity is high. Cluster similarity[7] is measured in regard to the mean value of the objects in a cluster, which can be viewed as a clusters centroid or centre of gravity. First, it randomly select K of objects, each of which initially represents a cluster mean or centre[6]. For each of the remaining objects, an object is used assigned to the cluster to which it is the most similar based on the distance between the objects and cluster mean. It then computes the new mean of each cluster. This process iterates until the criterion function coverage. Typically the square error criterion is used defined as

$$E = \sum_{i=1}^{k} \sum_{p \varepsilon C_i} |p - m_i|^2$$

Where **E** is the sum of square error of all the objects in the data set; **P** is point in space representing in the given object and **mi** is the mean of cluster $_{Ci}$, In other words, for each object in each cluster, the distance from the object to its cluster centre is squared ,and the distance are summed. The criterion tries to make the resulting K clusters as compact and as separate as possible.

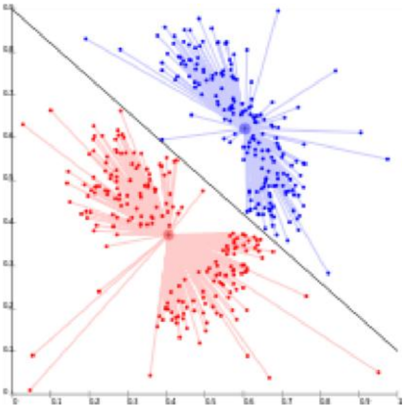

Fig.1.2 Grouping Of Clusters.

Algorithm: k-means for grouping of similar words

Input:

K: the number of clusters,

D: a dataset containing n objects,

Output: a set of k clusters,

Method:

1. Read the k cluster value from user;
2. **Repeat**
3. Find the representatives;
4. Assigning the cluster using the nearest value
5. Allocating the object to the representatives;
6. **Until** no change;

4.Conceptual Ontological Graph

Conceptual Graphs (CG) is a logical formalism that includes classes, relations, individuals and quantifiers. The main feature is standardized graphical representation that like inthe case of semantic networks allows human to get quick overview of what the graph means Conceptual graph[10] is a bipartite orientated graph where instances of concepts are displayed as rectangle and conceptual relations are displayed as ellipse.

Oriented edges then link these vertices and denote the existence and orientation of relation. It is represented as conceptual graph G=(C,R) where the concepts of the sentence are represented as vertices (C).The relations among the concepts such as agents, objects, and actions are represented as (R). C is a set of nodes (c_1,c_2,…..c_n) where each node C represents represents a concept in the sentence or a nested conceptual graph *G*; and *R* is a set of edges (r_1; r_2,….,r_m), such that each edge *r* is the relation between an ordered pair of nodes (c_i,.., c_j). The output of the role labelling task, which are verbs and their arguments are presented as concepts with relations in the COG representation. This allows the use of more informative concept matching at the sentence-level and the document-level rather than individual word matching. The concept-based model proposes new weight to each position in the COG representation to achieve more accurate analysis with respect to the sentence semantics. Thus, each concept in the COG representation[10] is assigned a proposed weight, which is *weight COG*, based on its position in the representation. The proposed *weightCOG*i assigned to each concept presented in the COG representation and is calculated by:

$$\mathbf{weightCOGi= tfweight_i * LCOG_i (4)}$$

In equation (2), the *tfweight*$_i$value presents the weight of concept *i* in document *d* at the document-level as shown in equation (2). The *LCOG*$_{i\text{value}}$ presents the importance of the concept *I* in the document *d* at the sentence-level based on the contribution of concept *i* to the semantics of the sentences represented by the levels of the COG representation. Consider the below example of COG:

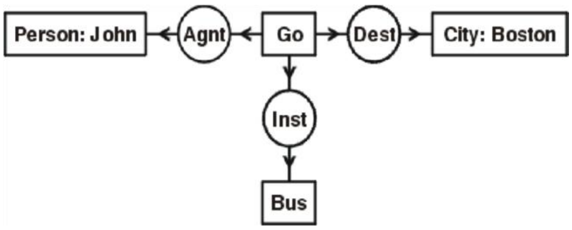

Fig.1.3 Relationship Between Two Objects.

5.Rapid Miner Tool

Rapid miner tools is used to data loading and transformation (Extract, transform, load) data preprocessing and visualization,modeling, evaluation, and deployment. Rapid Miner iswritten in theJava programming language. It uses learning schemes and attribute evaluators from the Weka machine learning environment and statistical modeling schemes from R-Project.The process-view of Rapid Miner offers a modular and pipelined view on the processed experiment. The process normally consists of four stages:

– the input stage,

– the pre-processing stage,

– the learning stage,

– the evaluation stage.

Text Input and tokenization

Texts or documents are available in various forms like plain ascii-files, xml- or html-files and pdf-files for example. Some text-formats are easy to process and some need additional parsing effort. The plugin offers an input-operator which can handle ascii-, html- and pdf-files. After loading a document, it is present as continuous block of text in one table cell, which is not accessible in a profitable way. As a first step tokenizers have to be used to split up the text. Up to now, a trivial sentence and word- tokenizer is available. While tokenizing splits the texts into smaller parts the original structure is still kept available. So, if one wants to process documents on the word-level, the sentencelevel (the level above) is still present and can be used for processing the features. These features basically can be extracted from the current position of thedocument (sentence, word ...) and from circumfluent positions.

Fig 1.4 Possible IE ExampleSet in Rapid Miner.

Token No	Do. NO	Sent . No	Token No	Att1	label
…….	......	….	…..	……	……
Semantic	2	4	2	“S“	O
Taxonomy	2	4	3	“T“	O
Rapid	2	4	4	“R“	LOC
Hypotheses	2	4	5	“H“	O
….	…..	……	……	…….	…..

Fig 1.5 Possible IE with attribute2 in Rapid Miner.

Token No	Doc. NO	Sent . No	Token No	Att2	label
…….	......	….	…..	……	……
Semantic	2	4	2	“c“	O
Taxonomy	2	4	3	“y“	O
Rapid	2	4	4	“d“	LOC
Hypothesis	2	4	5	“s“	O
….	…..	……	……	…….	…..

Annotation and visualization

A visualization operator allows to view the documents and to annotate parts of them. One can select attributes which shall be visualized, in order to highlight different aspects of the text. Figure 4 shows a document with some annotated words.

Pre-processing

The pre-processing-operators are used to enrich the document and its tokens with syntactical information which will later be used to extract semantic information. Tokens can be enriched with contextual information as well as they can deliver inner-token information. One should keep in mind that different tasks need different preprocessing. To use machine learning algorithms for IE, one has to enrich the documents, sentences or words with attributes extracted from themselves and from their context. The plugin offerspreprocessing-operators which can be used in a very modular way.

Learning

The Information Extraction offersoperators for NER and for RE. Rapid Miner learners deliver so called models which equate to a function and can be applied to an exampleset. Until now, for NER, the Conditional Random Fields (CRF) operator can be used. For RE the tree kernel should be used. The underlying techniques have been described before. The implementation of these learning algorithms has been kept modular to allow the combination of various methods. Due to this modularization the CRF-learner can be combined with various optimization-methods such as quasi-newtonmethods or evolutionary algorithms.

DM-usage the calculated model can of course be used for the extraction of interesting information from unseen documents, but after having processed entities or relations or every other information one can easily build up a new exampleset containing the extracted information to gain additional knowledge.

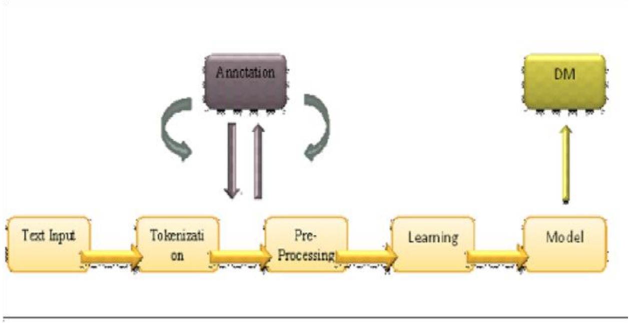

Fig.1.6 Information Extraction Stages.

Conclusion

Knowledge based NLP systems will be practical for real-world applications only when their domain-dependent dictionaries can be constructed automatically. Our approach to automated dictionary construction is a significant step toward making information extraction systems scalable and portable to new domains. In this paper we studied one important aspect of building a high quality information extraction system, that of refining dictionaries used in the extractor. We providing rigorous theoretical analysis and experimental evaluation of the optimization problems that such refinement entails. we have proposed a text mining mechanism which extracts entity relationships words from documents. Summarizing a document through relations help in easy visualization of contents of a repository.

For syntactic tagging, every word or phrase must be tagged whereas, for AutoScap, only the targeted information needs to be tagged. Sentences, paragraphs, and even texts that are irrelevant to the domain can be effectively ignored. We have demonstrated that automated dictionary construction is a viable alternative to manual knowledge engineering.

In previous explanations, AutoScap produces a concept node dictionary for the domain that achieved 100% of the performance of dictionary. On the other hand, AutoScap is critically dependent on a training corpus of texts and targeted information. This is an attractive idea since it can help in direct future research and provide interesting insights into a domain.

•

References

1. W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
2. Ellen Riloff Department of Computer Science University of Massachusetts, Proceedings of the Eleventh National Conference on "Artificial Intelligence", 1993, AAAI Press / MIT Press, pages 811–816.
3. R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In Proceedings of First International Conference on "Knowledge Discovery and Data Mining", pages 112 - 117, 1995.
4. W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.
5. S. Shehata, F. Karray, and M. Kamel. "Enhancing text clustering using concept- based mining model". In ICDM, pages {1043,1048}, 2006.
6. Y. Yang, "An Evaluation of Statistical Approaches to Text Categorization Information Retrieval", vol. 1, pp. 69-90, 1999.
7. Ning Zhong, Yuefeng Li, and ShengTang Wu, "Effective Pattern Discovery for Text Mining", IEEE transactions on knowledge and data engineering, vol. 24, no. 1, january 2012.
8. LipikaDey, Muhammad Abulaish, Jahiruddin and Gaurav Sharma, "Text Mining through Entity-Relationship Based Information Extraction", 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops.
9. Proceedings of the Eleventh National Conference on Artificial Intelligence, "Automatically Constructing a Dictionary for Information Extraction Tasks" by Ellen Riloff, 1993, AAAI Press / MIT Press, pages 811–816