Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20240301011
Volume 3, Issue 1 (January-April 2024), PP: 74-77
www.indjcst.com
ISSN No: 2583-5300

A Review- Machine Learning Techniques for Text Summarization

Sivakumar Nagarajan
Technical Architect, I & I Software Inc, 2571 Baglyos Circle, Suite B-32, Bethlehem, PA-18020, USA.

Abstract: *Aspect ranking framework is significant to identify the important aspects from numerous consumer reviews posted in various domains like hotel, movie and product etc. They could be broadly classified into supervised and unsupervised approaches broadly. Supervised methods rely on semantic knowledge bases. These are found to be effective for ranking compared to conventional approaches. These methods available in the literature are discussed in detail. Next, this review focuses on the extractive summarization systems, in which the summary is generated by picking a sub-set of sentences from the related text. Extractive summarization systems that utilize machine learning, optimization and map reduce framework are explained elaborately. This is due to the efficiency of these techniques reported in the comprehensive works available for text summarization. A literature covering text similarity discovery methods employing text, semantic information and graph based systems are presented in detail at the end of the chapter. Among these graph based methods play a vital role in current field of the research.*

I.INTRODUCTION

Aspects are significant features in customer reviews that are essential for analyzing the reviews to serve in various business decisions. These reviews are often unstructured and do not imply any meaning in the text. Summarizing the core characteristics of these aspects becomes important for commercial purposes. Both aspect ranking and summarization had been accomplished by different machine learning techniques. This research focuses on utilizing customer preferences for ranking the aspects from the reviews in order to improve the business decisions from the stake holders. This facilitates to improve aspect ranking from the customer's point of view. Summarization had been explored using machine learning and optimization with parallel and largescale analytic strategies. This enables to process large volume of customer reviews and improve the quality of the summary generated for the aspects. Employing machine learning and parallel algorithms will enable to improve the quality of text summarization systems. Current researches in the area of text mining deals the problems of text representation, classification, clustering, text summarization and modeling of hidden patterns.

Text mining is an area where large amount of unstructured text is analyzed to gain some actionable intuitions. Natural language processing and text mining could be viewed as artificial intelligence technologies to enable transforming key contents from large text to quantitative information. This could be used for further analysis and would help in business processes. Alternate terms for text mining are text data mining and text analytics. In the present age, unstructured text is found in huge volume with internet as key source of data.

Most of these unstructured data is generated from millions of customer reviews. Organizations would be able to make better decisions when these reviews are quantitatively analyzed. Identification of key features and summarizing the key content into meaningful form are the two thriving factors for the wide application of text mining. These are accomplished by feature extraction and text summarization. When opinions in the text were added to these tasks, they gain deep insights in to operational challenges faced by the prospective customers.

II.LITERATURE REVIEW OF TECHNIQUES FOR TEXT SUMMARIZATION

Aspect ranking framework is significant to identify the important aspects from numerous consumer reviews posted in various domains like hotel, movie and product etc. They could be broadly classified into supervised and unsupervised approaches broadly. Supervised methods rely on semantic knowledge bases. These are found to be effective for ranking compared to conventional approaches. These methods available in the literature are discussed in detail. Next, this review focuses on the extractive summarization systems, in which the summary is generated by picking a sub-set of sentences from the related text. Extractive summarization systems that utilize machine learning, optimization and map reduce framework are explained elaborately. This is due to the efficiency of these techniques reported in the comprehensive works available for text summarization. A literature covering text similarity discovery methods employing text, semantic information and graph based systems are presented in detail at the end of the chapter. Among these graph based methods

Aspect Ranking

Aspect identification and ranking became a core task in the field of text analytics. Identification of important aspects in a review (. becomes necessary as both consumers and firms need analytic reports for effective decisions. End-users pay attention to the important aspects. Firms concentrate on improving the quality and their business reputation which can be used in industry.

Aspect ranking systems could be categorized based on the information they use for ranking. The related approaches with literature relevance are stated below:

Term weight based Ranking Approaches

Term weight approaches are based on the heuristics that depend only on the frequency of features into account for finding most important features. These methods are quite useful when terms appearing in the unstructured text of documents are to be considered. Some systems retrieve opinion words by finding the frequent feature adjective word. These systems are simple in nature and easily retrieve frequency of the adjective words by using lexicons or predefined set of words. But, they are inefficient in finding semantically equivalent terms and depend on only adjective words.

Sentiment Analysis based Ranking Approaches

Aspect based opinion mining or sentiment analysis deals with extracting and selecting the features to give a summary regarding the opinions expressed about the feature. Utilization of association rule mining for identifying frequent features (Hu & Liu 2004) focus on mining opinion/product features that the reviewers have commented on and mining them to generate an opinionated summary. OPINE system was developed (Maria & E-zine 2005) based on sentiment labels with relaxation labeling. Keyword matching strategy was used (Zhang *et al.* 2010) to identify the essential features. This limits the features only based on keywords. An approach based on sentiment patterns was deployed by (Zhail *et al.* 2010) and depends on the structural characteristics of reviews. This approach does not rely on the intended semantics of an aspect.

Semantic Information based Ranking Approaches

Semantic information systems try to exploit semantic facts using knowledge base or ontology. Ontology is the study of existence. It is also the study of how we determine if things exist or not, as well as the classification of existence. It attempts to take things that are abstract and establish that they are, in fact, real. Technically it denotes an artifact (Bloehdorn *et al.* 2004) that is designed for a purpose, which enables the modeling of knowledge about some domain, real or imagined. Knowledge base is an organized collection of facts about the system's domain. Inference mechanisms are required to interpret the facts from the knowledge base. Limited systems using simple knowledge bases (Buche *et al.* 2013) are available for aspect ranking. They are not explored much because of its complexity.

III.ASPECT BASED TEXT SUMMARIZATION

Aspect based text summarization is an essential task in the field of text mining when large volume of text data is to be analyzed. Today's web contains numerous reviews and this makes stakeholders difficult to go through the entire content. But this content becomes essential when operational decisions have to be made by them. These decisions are crucial for both firms and its stakeholders to improve quality in their business and to sustain reputation for the industry. An aspect based summarization system takes as input a set of user reviews for a specific product or service and produces a set of relevant aspects, an aggregate score for each aspect and supporting textual evidence. Aspect based text summarization systems can be classified based on the techniques used for summarizing the content.

Machine Learning Approaches

Clustering had been widely used for text summarization assignment, among the various machine learning approaches available. Ensemble method in clustering is a technique included in a hybrid collection of machine learning models in order to obtain improved performance than any model in the collection. K-means is a famous clustering algorithm, which is one of the simplest unsupervised machine learning algorithms, and it is usually very fast. Bagging is one of the most popular ensemble techniques. K-means is also one of the top-10 algorithms in data mining.

However, there is no guarantee that it will converge to the global optimum, it is a heuristic algorithm, and different choices of initial clusters may produce different results. In recent years, cluster analysis has become significant in text mining. There are several variation algorithms of k-means: Meanwhile, there remain several problems of those variation algorithms. An important problem of wide concern is k-mean's instability and sensitivity to outliers. Map Reduce is programming model and an associated implementation for processing and generating large data sets for parallel algorithms. Parallelk means clustering algorithm had been a focus for researchers to decrease the time complexity in clustering large datasets. Aspect based summaries for review texts needs improvement of the clustering algorithm involving more appropriate decision of representative sentences. Some of the problems identified in k means clustering based on the survey are instability and accuracy of the algorithm. It is also found to be more sensitive to outliers when used for text summarization problem. Hence there is a need for optimization based approaches as presented below.

Optimization Approaches

Substantial systems used for review summarization are described in this section. Opinos is, a recent tool for generating aspect based summaries, is an unsupervised graph optimization based approach, achieved improved performance but it cannot group sentences at a deep semantic level. Some systems which use optimized domain knowledge, typically make the highly limiting assumptions that no prior knowledge of the domain being summarized is available. Involving sentiment classification makes feature selection as an important phase in summarization. Feature selection approaches using single feature ranking and Binary Particle Swarm Optimization (BPSO) had also been found in the literature. In the single feature ranking approach, only significant top features are ranked.

Multi objective approaches

There are many multi objective approaches in the literature, where PSO is significant in the process of feature selection, extraction and text summarization. This section discusses about the multi objective systems and their applications in text summarization. Incorporating NSGA-II into PSO had a vital impact in the literature using multi objective optimization (Deb *et al.* 2002). A sigma method for multi objective optimization in which best local guides are adopted to improve the convergence and diversity of PSO had been projected (Mostaghim & Teich 2003). In another work Li (2003 & 2004) projected max-min PSO, which uses a fitness function that requires no additional clustering or niching procedure to maintain diversity.

Map Reduce Approaches

Detailed literature studies about text summarization approaches using map reduce are studied in this section. Aspect-based sentiment summarization has been considered extensively in the literature in many systems. Map Reduce is a parallel programming model used for analyzing large data sets. Machine learning techniques like parallel k means clustering algorithm had been extensively used by researchers to reduce the time complexity in clustering large datasets. Some of the recent works for aspect summarization using Map Reduce include machine learning techniques like Support Vector Machine (SVM) and two different Map Reduce stages for aspect summarization. Furthermore, dealing with high dimensions and large data sets can be problematic due to inefficiency.

IV.TEXT SIMILARITY IDENTIFICATION

This section presents a study on various techniques used in detection and computation of document similarity. Text based, graph based approach and semantic similarity metrics based measures have been utilized by various researchers. Many Approaches measure text similarity statistically based on the word distribution. They measure the similarity assuming that words occurring in the same context tend to be similar in meaning. By inferring semantics from text without using explicit knowledge, word-level approaches become susceptible to many problems.

Text based approaches

Some of the commonly used text based techniques in document similarity identification and computation are string based and semantic relation based approaches. An algorithm is employed to associate noun phrases in an aligned bilingual corpus. The taggers provide part- of speech categories which are used by finite-state recognizers to extract simple noun phrases for both languages. Noun phrases are then mapped to each other using an iterative re-estimation algorithm that bears similarities to the Baum- Welch algorithm which is used for training the taggers. The algorithm provides an alternative to other approaches to find word correspondences, with the advantage that linguistic structure is incorporated. Improvements to the basic algorithm are described, which enable context to be accounted for when constructing the noun phrase mappings.

Semantic Approaches

Semantic methods are commonly dependent on the measures used for similarity computation. A new measure of semantic similarity using 'is-a' taxonomy has been presented, which is based on the notion of information content. State of art semantic similarity computation measures (Lingling Meng *et al.* 2013) had been studied including path, information and hybrid approaches. The path-based measure computes similarity between two concepts as a function of the length of the path linking the concepts and the position of the concepts in the taxonomy. Semantic approaches to match relevant web pages in terms of topical intent for contextual advertising had been used (Broder *et al.* 2007 and Woo-Jong, Ryu & Jung-Hyun Lee 2013). In this work, it seeks to utilize the verbal intent, which complements topical intent, in semantic contextual advertising.

Graph Based Approaches

Stochastic graph-based method (Gunes Erkan *et al.* 2004) had been employed for computing relative importance of textual units and a detailed analysis of Lex rank approach and applied it to a larger data set. They discuss how random walks on sentence-based graphs can help in text summarization and also briefly discuss how similar techniques can be applied to other natural language processing (NLP) tasks such as named entity classification, prepositional phrase attachment, and text classification. Graph-based centrality has several advantages over Centroid. First it accounts for information subsumption among sentences. If the information content of a sentence subsumes another sentence in a cluster, it is naturally preferred to include the one that contains more information in the summary. Analysis of document similarity can be calculated efficiently compared to other graph-traversal based approaches by using similarity measures (Christian Paul *et al.* 2016). The measures should provide a significantly higher correlation with human notions of document similarity. Their approach holds good for short documents with few annotations. During similarity calculation, query document is expanded and full search is performed. Several approaches for graph based document were proposed based on the nlexical knowledge graph Word Net. This achieves a highly scalable solution by performing all knowledge graph related work in the Semantic Document Expansion pre-processing step. The task of computing document similarity has been studied using semantic similarity. Semantic similarity between concepts in knowledge graphs (KGs) such as WordNet and DB pedia are measured using wpath. This combines information content to weight the shortest path length between concepts.

V.CONCLUSION

A detailed survey of work related to feature ranking methods, summarization techniques and text similarity identification mechanisms has been carried out. With respect to feature ranking, semantic approaches register good performance. Domain knowledge and author preferences have been less explored in semantic methods that have been used extensively in feature

ranking systems.

Hence author preference and ontology for aspect ranking system was used. The survey on text summarization indicates that clustering is best suited for text summarization. Existing methods include partitioning based clustering algorithm, but suffers from instability. This has inspired the proposal of parallel clustering with bagging to overcome system instability thereby improving the quality of the summary for better performance. Study on optimization techniques leads the extensive use of single objective optimization for text summarization. Design of Multi objective functions including optimization had been found minimal for text analytics applications. This has influenced the conception of feature summarization system with Multi objective optimization using PSO encompassing two different text representation models Recent findings suggest the use of Map reduce model for large scale text summarization. There have been fewer explorations found to apply parallel algorithms with map reduce for summarization. This paved way for construction of an in node optimizer algorithm using mapper and combiner for feature summarization.

Reference

1. *Fattah, M.A. A hybrid machine learning model for multi-document summarization. Appl. Intell.* ***2014****, 40, 592–600.*
2. *Zhong, S.; Liu, Y.; Li, B.; Long, J. Queryoriented unsupervised multi-document summarization via deep learning model. Expert Syst. Appl.* ***2015****, 42, 8146–8155.*
3. *Yao, C.; Shen, J.; Chen, G. Automatic document summarization via deep neural networks. In Proceedings of the 2015 8th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 12–13 December 2015; Volume 1, pp. 291–296.*
4. *Singh, S.P.; Kumar, A.; Mangal, A.; Singhal, S. Bilingual automatic text summarization using unsupervised deep learning. In Proceedings of the 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, India, 3–5 March 2016; pp. 1195–1200.*
5. *Yousefi-Azar, M.; Hamey, L. Text summarization using unsupervised deep learning. Expert Syst. Appl.* ***2017****, 68, 93–105.*
6. *Chopade, H.A.; Narvekar, M. Hybrid auto text summarization using deep neural network and fuzzy logic system. In Proceedings of the 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, India, 23–24 November 2017; pp. 52–56.*
7. *Shirwandkar, N.S.; Kulkarni, S. Extractive text summarization using deep learning. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018; pp. 1–5.*