IMPACT OF THE MACHINE LEARNING FEATURE ON THE DEVELOPMENT OF TECHNOLOGY

[1] Ravindran R [2]Dr. Rajeev Yadav
[1]Research Scholar, [2]Supervisor
[1-2] Department of Computer Science and Engineering, OPJS University, Distt. Churu, Rajasthan, India

Accepted: 01.07.2021 **Published**: 01.08.2021

ABSTRACT

"A total of six billion people reside in developing nations. Researchers in subjects ranging from sociology to statistics, ecology to economics have long been interested in the specific issues that these places confront. With the rapid maturation of machine learning (ML) approaches, academics are increasingly turning to machine learning to enhance development research and practice. To provide an example, supervised machine learning methods may be used to offer expert decision assistance for health care in resource-limited areas, while deep learning techniques can be used to analyze satellite data to develop fresh economic indicators. Nonetheless, there are significant obstacles to deploying machine learning in the underdeveloped world. Data availability, processing capability, and Internet connectivity in developing nations are often much less developed than in developed ones. The convergence of machine learning's enormous promise with the actual constraints faced by developing world environments has sparked a growing corpus of study in machine learning for developing world environments (ML4D). In this essay, we will take a look at this increasing body of literature. For the purpose of determining the scope of the study, we provide a formal definition for ML4D and conduct a survey of significant application issues in the field. For the design and implementation of ML4D initiatives that promote important development goals, we provide best practices derived from the literature. Throughout this essay, we will concentrate on research that heavily depends on machine learning. Since a result, we will refrain from discussing themes that are exclusively related to Big Data or information technology, as they have already been well covered in prior publications".

These characteristics are applicable to both the issues which are being discussed and the remedies which are being suggested to those challenges.

It takes into consideration five principles of growth: "social, economic, health, environmental, and institutional development", to name a few.

Keywords:- Development, Significant and Research.

INTRODUCTION

"One effective way of looking at machine learning is to think of it as a process of searching through a very vast space of alternative hypotheses in order to find the one that best matches the observed data and any previous information that the learner has. Consider, for example, the set of hypotheses that may theoretically be produced by the checkers learner described above. This hypothesis space consists of all evaluation functions that may be represented by a particular set of values for the weights wo through w6, and it is composed of all possible combinations of these values. In order to find the hypothesis that is most consistent with the training examples supplied, the learner must search through a large amount of information. It accomplishes this purpose by repeatedly tweaking the weights, adding a correction to each weight each time the hypothesized evaluation function predicts a value that is different from the training value, as seen in Figure 1. Specifically, when the hypothesis representation examined by the learner specifies a continuously parameterized space of alternative hypotheses, this technique performs well. These algorithms seek a hypothesis space specified by some underlying representation, and they are shown here (e.g., linear functions, logical descriptions, decision trees, artificial neural networks). These various hypothesis representations are ideal for learning various types of target functions, as indicated by their names. In order to arrange the search through the hypothesis space, the appropriate learning algorithm for each of these hypothesis representations uses a different underlying structure to organize the search through the hypothesis space. The viewpoint of learning as a search issue will be revisited throughout this book in order to describe learning techniques by their search strategies and the underlying structure of the search spaces that they investigate. Furthermore, we will find this point of view useful in formalizing our analysis of relationships such as those between the size of the hypothesis space to be searched, the number of training examples available, and the level of confidence we can have that a hypothesis consistent with the training data will correctly generalize to unseen examples".

APPLICATIONS OF ML TO THE SEMANTIC WEB

"The interest in implementing machine learning methods in the context of the Semantic Web has grown in recent years, and particular machine learning tracks and workshops have been established at the major Semantic Web conferences. Ontology learning is one of the applications of machine learning algorithms that is

covered. Ontologies may be learned from scratch using machine learning methods, and they can also be used to improve previously existing ontologies. Learning data comes from a variety of sources, including Linked Data, social networks, tags, and textual data. The learning of the mapping from one ontology to another (e.g., based on association rules or similarity-based approaches) is another common use of machine learning. Some suggested ways for learning from the Semantic Web are based on Inductive Learning Processes (ILP) (e.g., classification or association learning). Such an approach is supplemented by newly created tools for ontology-based data mining, such as DL-Learner, RMonto, and SDM-Toolkit, which are all available for download. A particularly intriguing use of this kind was developed as part of the e-LICO project. It consisted in optimizing knowledge discovery processes through ontology-based meta-learning, which is machine learning from meta data of previously executed experiments, where meta data was represented with background ontologies. This represents a perspective of ILP for the Semantic Web, in that it is machine learning from meta data of previously executed experiments. Ontology learning and ILP are based on the assumption of deterministic or nearly deterministic relationships. The growth in interest in machine learning methods is primarily owing to the open, dispersed, and intrinsically imperfect character of the Semantic Web, which has sparked this interest. In such a setting, using solely deductive strategies, which have typically dominated reasoning approaches for ontological data, becomes difficult. As part of the LarKC project, a scalable machine learning technique has been created that is effective in dealing with the high-dimensional, sparse, and noisy data that is encountered in those fields of application. Matrix factorization is used in this method, and it has shown improved performance on a number of Semantic Web data sets so far. Extensions have been created that may account for temporal effects, model sequences, and add ontological background and textual information, among other things. In the ISWC Semantic Web Challenge, the technique was part of the winning submission, which was based on it.

REVIEW OF LITERATURE

Despite the many ideas, the actual quality of QBIC systems has a great history to go before it achieves the precision that has been sought. Despite the fact that numerous corporate CBIR have been built, a mature information assess alternative is not yet produced; nonetheless, this sector is actively evolving. These findings were drawn from a study conducted by the authors:

- As a result of the literature study, it is clear that the all image pattern recognition studies strive for high accuracy while also developing algorithms that are simple to implement (essential for real world applications) and that create only even before.
- Per the papers analysed, it seems that machine learning techniques are now the most common approaches for image representation at the present time. Furthermore, it seems that the bad and excellent performances of each method are largely reliant on the feature set that describes the query as well as database pictures, according to the results. In other aspects, the effectiveness of the CBIR system is determined by the type of the characteristics that are used.
- Visual characteristics may be divided into three categories: elementary characteristics such as colour or form, logical characteristics such as the identification of things displayed, and abstract characteristics such as the importance of scenes shown. Nonetheless, all presently available methods only make use of rudimentary traits unless human annotation is used in conjunction with the visual characteristics. As a result, in order to enhance the effectiveness of CBIR systems, it is required to improve the image processing techniques.
- Picture retrieval practitioners often use Classification Technique (PCA) to deal with large-scaled images based and to deal with the 'high dimensional,' which is a term that refers to the phenomenon of having too many variables.
- In order to reduce the temporal complication of the Cnn models, several of them use an indexing technique. In most cases, the technologies are either shrub or cluster-based in nature, with each holding its own set of pros and drawbacks. A approach that incorporates both of these elements will be a fascinating topic of investigation.

ASPECT RANKING

Aspect identification and ranking became a core task in the field of text analytics. In order to make successful judgments, it is crucial to identify significant components in an assessment (That such et ill. 2013). This is because both customers and businesses want analytic reports. Most final pay heed to the most critical features. Firms focus on enhancing the quality of their products and the prestige of their company, which may be applied in the manufacturing sector. 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 techniques that take simply the occurrence of characteristics into consideration (Hu & Wang 2005) for said purpose of identifying the most significant features When this comes to analysing phrases that exist in unstructured information of files, these strategies are quite beneficial. One method used by certain

computers (Lin et al. 2014) to obtain opinion mining is to look for the most often occurring feature description word.

Sentiment Analysis Based Ranking Approaches

Essential piece opinion mining, also known as trend analysis, seems to be the process of extracting and choosing characteristics in order to provide a synopsis of the views stated about a certain aspect. Use of mining to find frequent characteristics (Hu and Xiao 2004) focuses on mining viewpoint elements that have been remarked on by users and on which they have agreed or disagreed. mining them to generate an opinionated summary.

Semantic Information Based Ranking Approaches

Semantic information systems try to exploit semantic facts using knowledge base or ontologies. Inquiry of the nature of what it means to be alive. It is really the investigation of how humans judge whether or not something exists, and also the categorization of things that exist. It makes an effort to prove the reality of abstract concepts by establishing that they really are, in truth, 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.

EXPERIMENTALRESULTS

DatasetDescription

The trials are now underway and implemented using Java in 8GB RAM with dual-core processor. The data set considers.

Set Of Data Of the Movie Industry

Unsupervised classification can be performed upon that dataset, which contains 50,000 unlabeled papers. The data source has a maximum volume of 400 Kilobytes in total.

Data From the Hotel Review

- Detailed evaluations of hoteliers in ten multiple towns
- Within every city, there seem to be approximately 80-700 hotels.
- The deadline, the title of the overview, and the comprehensive review are among the field that have been extracted
- A average of 259,000[190 Megabits per second] evaluations have been submitted.
- There are ten different folders, each representing one of the ten cities mentioned
- Previously. Each document (inside each of these ten folders) would comprise all of the
- review sites pertaining to a specific hotel.
- The results of each stage with comparative performance measures are presented below

Sample data with details is given inTable3.1.

Table -3.1 Data set Statistics-Movie Domain

Movie Name	Review#	Sentence#
2 States	200	700
Penguins of Madagascar	150	850
Happy Film	200	800
The Dark Knight	100	700
Pulp Fiction	150	650
The Dark Knight	340	550
The Godfather	320	460

PROPOSED ASPECT SUMMARIZATION USING PARTICLE SWARM BASED MULTI-OBJECTIVE OPTIMIZATION

Particle swarm optimization (PSO) has been widely adopted for various applications. It has been extended to use in other fields of optimization like constrained The terms optimizing, inter optimization, and so on, are all used interchangeably. The PSO optimizer, when used for decision variables optimization, must take into account Pareto domination each time it refreshes components and retain non-dominated solutions to approach the Pareto front as closely as possible. On the playing field of text summarization, PSO had been studied. A diversity enhanced PSO algorithm for text summarization has been developed. A new multi objective fitness function has been used in PSO with discrete and continuous variations. The multi objective function used tries to retain the sentences which are swarms with high score and minimum similarity. This avoids redundant sentences to be included in the summary, which improves the overall quality of the feature based summary generated. The proposed system adopts PSO for review summarization using multi objective functions and explained in the following section.

1) Flow Diagram of PSO based Text Summarization System

PSO based text summarization system employs multi objective functions with different text representations. Input reviews are pre-processed and represented with two different representation using bag of words (BoW) and real valued vector models. Then PSO based text summarization algorithm is executed with the pre-processed reviews and summary is generated.

PROPOSED ASPECT SUMMARIZATION USING PARTICLE SWARM BASED MULTI-OBJECTIVE OPTIMIZATION

Particle swarm optimization (PSO) has been widely adopted for various applications. It has been extended to use in other fields of optimization like constrained The terms optimizing, inter optimization, and so on, are all used interchangeably. The PSO optimizer, when used for decision variables optimization, must take into account Pareto domination each time it refreshes components and retain non-dominated solutions to approach the Pareto front as closely as possible. On the playing field of text summarization, PSO had been studied. A diversity enhanced PSO algorithm for text summarization has been developed. A new multi objective fitness function has been used in PSO with discrete and continuous variations. The multi objective function used tries to retain the sentences which are swarms with high score and minimum similarity. This avoids redundant sentences to be included in the summary, which improves the overall quality of the feature based summary generated. The proposed system adopts PSO for review summarization using multi objective functions and explained in the following section.

1) **Flow Diagram of PSO based Text Summarization System**

PSO based text summarization system employs multi objective functions with different text representations. Input reviews are pre-processed and represented with two different representation using bag of words (BoW) and real valued vector models. Then PSO based text summarization algorithm is executed with the pre-processed reviews and summary is generated.

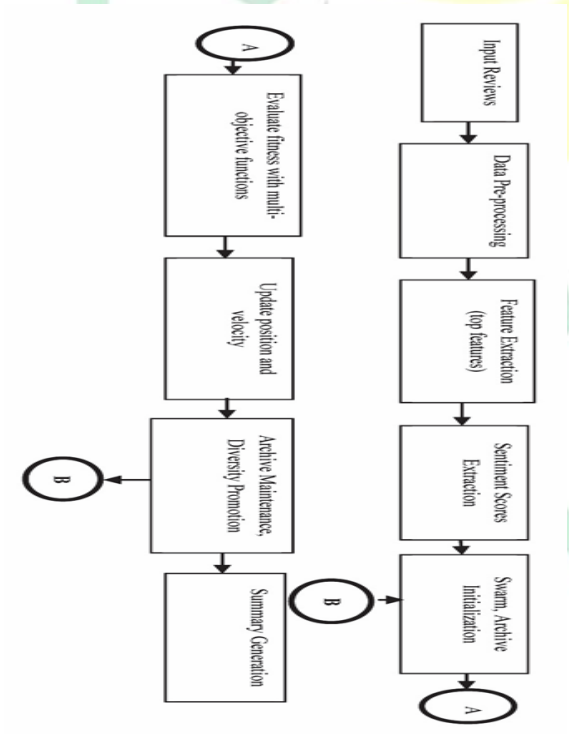

Fig. 4.1 Flow Diagram of the Proposed PSO based Framework for SurveyRundown

RESULT COMPARISON USING STATISTICAL TEST (WILCOXON SIGNED RANK TEST)

Distributed Clustering with Optimization Method

The table 4.1 shows the Wilcoxon signed rank test for hotel domain features. The algorithm B indicates the parallel k means clustering algorithm. The proposed algorithm A denotes distributed clustering algorithm with optimization.

Table - 4.1 **RESULT COMPARISON USING STATISTICAL TEST (WILCOXON SIGNED RANK TEST)**

S. No.	Feature	parallel k means clustering algorithm B	Distributed clustering algorithm with optimization algorithm A	Diff (A-B)	Ordered diff	Rank	Signed Rank
1	Location	0.64	0.68	0.04	0.02	1	1
2	Food	0.61	0.7	0.09	0.03	2	2
3	Service	0.48	0.5	0.02	0.03	3	-3
4	Rooms	0.68	0.71	0.03	0.04	4	4
5	Staff	0.36	0.48	0.12	0.05	5	-5
6	Price	0.56	0.51	-0.05	0.05	6	6
7	Facility	0.51	0.56	0.05	0.09	7	7
8	Comfort	0.49	0.59	0.1	0.1	8	8
9	Rating	0.52	0.49	-0.03	0.12	9	9
						W-	-8
						W+	37
						W	37

The same test is applied for movie domain data set and the results using ROUGE 1 with f measure scores are shown in Table 4.1.

Table - 4.2
Statistical Test Performance for Distributed Clustering Method from Movie Domain

S. No	Feature	parallel k means clustering algorithm B	Distributed clustering algorithm with optimization algorithm A	Difference (A-B)	Ordered difference	Rank	Signed Rank
1	Actor	0.56	0.59	0.03	0.02	1	1
2	Story	0.61	0.67	0.06	0.03	2	2
3	Screenplay	0.47	0.49	0.02	0.05	3	3
4	Direction	0.58	0.7	0.12	0.06	4	4
5	Dialogue	0.42	0.47	0.05	0.07	5	5
6	Cinematography	0.39	0.48	0.09	0.08	6	6
7	Music	0.4	0.48	0.08	0.09	7	7
8	Editing	0.45	0.52	0.07	0.1	8	8
9	Performance	0.56	0.66	0.1	0.12	9	9
						W+	45
						W	45

In the movie domain data set, summaries of all the features considered are showing positive ranks using signed rank test. The test statistic also was found to be 45. This shows enhanced performance of the proposed algorithm with existing algorithm. Null hypothesis is rejected with the significance level of 0.05.

CONCLUSION

The approach uses parallel clustering technique with the basic principle derived from k- means algorithm. This is deployed using Map Reduce. When map reduce is used it minimizes the inefficiency of K-means to process large scale and noisy data. Natural language based semantic features improved the quality of summary for the significant features identified. Other than using NLP features, text representation with optimization enhanced the overall performance of the synopsis and reduced the number of distracting phrases that would have been included in the synopsis

The next technique is designed with two different text representations, real valued vectors and Bag of words model. Bag of words model would map one sentence for one feature whereas real valued vector represented multiple features in a sentence. Particle swarm optimization algorithm using improved diversity had been developed and tested with these two representations. Continuous and discrete models were established and verified with customer reviews. Real valued continuous model with PSO had better performance in the quality of the summary. This is measured with ROUGE metrics and increased by 5 % compared to other existing systems.

The subsequent method for summarization involved in-node combiner optimization. A partitioner and an in-node mapper algorithm was developed and tried using Map reduce. The noisy sentences were removed from the reviews using optimization and also this minimized processing time substantially. The quality of the summary also displays an improvement around 6 % when using ROUGE metrics.

In the most significant component of the study, the production of a text has been undertaken. similarity assessment method using graph databases and grammatical knowledge. The text summaries generated by the above mentioned three summarization systems were used for similarity assessment. Involving grammatical linkages and graph databases for retrieval improved the efficiency of the similarity identification on an average of 4% with jaccard coefficient, cosine similarity and dice coefficient metrics. The system was also tested with other standard datasets like Microsoft research paraphrase corpus and had performed with effective outcomes.

REFERENCES

- Nyaung, DE &Thein, TLL 2015, 'Feature-based summarizing and ranking from customer reviews', World Acad. Sci. Eng. Technol. Int. J. Comput. Electr. Autom. Control Inf. Eng, vol. 9, Issue 3, pp. 734–739.
- Pang, B & Lee, L 2008, 'Opinion mining and sentiment analysis,' In Foundations and Trends® in Information Retrieval, vol. 2, Issue 1, pp. 1-135.
- Philip Resnik 1995, 'Using Information Content to evaluate semantic similarity in a Taxonomy', ACM digital library, pp. 18-25.
- Priya, V &Umamaheswari, K 2016, 'Ensemble based Parallel k means using Map Reduce for Aspect Based Summarization', In: International Conference on Informatics and Analytics Article no 26, Pondicherry, India.
- Ryosuke Tadano, Kazutaka Shimada & Tsutomu Endo 2010, 'Multi- aspects Review Summarization Based on Identification of Important Opinions and their Similarity', In Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation PACLIC 24, pp. 685-692.
- Ryu, WJ, Lee, JH & Lee, S 2017, 'Utilizing verbal intent in semantic contextual advertising', In: IEEE Intelligent Systems, vol. 32, Issue 3, pp. 7-13.
- SamanehMoghaddam& Martin Ester 2015, 'Aspect based opinion mining from Online Reviews', https: //www.cs.sfu.ca/~ester/ papers/SIGIR2012, Tutorial.Final.pdf, last accessed- 17 Apr, 2015.
- Sanchez-Gomez, JM et al. 2018, 'Extractive multi-document text summarization using a

multi-objective artificial bee colony optimization approach', In: Knowledge-Based Systems, vol. 159, pp. 1-8.

- Schuhmacher, M &Ponzetto, SP 2014, 'Knowledge-based graph
- document modeling', In: WSDM. pp. 543–542.
- Shah Neepa& Mahajan Sunitha 2014, 'Distributed Document Clustering Using K-Means', International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, pp. 24-29.
- ShefaliPatil, G & Bhatia, A 2014, 'Graph Databases-An Overview,' In: International Journal of Computer Science and Information Technologies, vol. 5, Issue 1, pp. 657-660.