VTKHI: Video Tagging Framework Using Knowledge-Driven Hybrid Intelligence and Quantitative Semantics Driven Reasoning

Gerard Deepak[1,†], Asritha Boddu[2,*,†]

[1]BMS Institute of Technology and Management,Bengaluru,India

[2]Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Bengaluru, India

Abstract

In Web 3.0 there is a need for a strategic framework for video tag recommendation that encompasses hybrid intelligence and is knowledge-centric. This paper proposes a video tag recommendation framework that incrementally enriches knowledge addition in the perspective of the dataset through topic modelling, using standard Semantic wikis, knowledge store repositories and a dynamic knowledge stack generation increasing the concentration of information in the proposed framework. Subsequently, the proposed framework hybridizes classification but independently identifies the stages by categorizing the dataset through the utilization of a CNN classifier and the generated dynamic knowledge stack classification is achieved by classification using the XGBoost machine learning classifier which preserves the computational complexity. The presence of quantitative Semantic relatedness is computed by using Simpson's diversity Index, Soccer league competition algorithm, SOC-PMI, as a criteria function that facilitates yielding the intermediate solution state which makes it quite efficacious. An overall precision of 96.74%, recall of 97.43%, accuracy of 97.08. F-measure of 97.08%, and FDR of 0.04% is achieved by the proposed VTKHI framework for video tag recommendation.

Keywords

Video Tagging, CNN Classifier, XGBoost, Soccer League Competition Algorithm, SOC-PMI, Simpson's Diversity Index,

1. Introduction

The proposed Video Tag Knowledge Hybrid Intelligence (VTKHI) framework is a pioneering solution designed for the challenges posed by Web 3.0 and the Semantic Web. Aligning with the principles of the Semantic Web, VTKHI integrates advanced topic modelling techniques and leverages Semantic wikis and knowledge store repositories to enhance its dataset dynamically. The framework adopts a hybrid intelligence strategy, combining Convolutional Neural Network (CNN) for dataset classification and XGBoost for dynamic knowledge stack classification, striking a balance between computational complexity and precision. Noteworthy is the incorporation of quantitative Semantic relatedness measures, such as Simpson's diversity index and Soccer league competition algorithm, aligning with the Semantic Web's goal of enriching information meaning.

MSW-2024: 4th International Workshop on Multilingual Semantic Web, December 10, 2024, Paris, France

*Corresponding author.

†These authors contributed equally.

gerard.deepak@bmsit.in (G. Deepak); aboddu2723@gmail.com (A. Boddu)

0000-0003-0466-2143 (G. Deepak); 0009-0003-9019-3722 (A. Boddu)

With exceptional performance metrics, including high precision, recall, and accuracy, VTKHI stands out as a cutting-edge solution poised to redefine standards for video tag recommendation in the evolving digital landscape of Web 3.0 and the Semantic Web.

1.1. Motivation

The primary motivation of the suggested framework is the lack of Semantically driven Web 3.0 compliant models of video tagging. Although the model exists they do not have a hybrid intelligence framework that encompasses the strategic amount of knowledge synthesized in several heterogeneous sources. Most importantly the quantitative reasoning through knowledge-centric Semantics which is the ultimate need to support the Web 3.0 paradigm which exhibits a high concentration compared to Web 2.0. It is not implemented in most of the existing frameworks of video tagging which is the primary focus of the proposed work.

1.2. Contributions

The major contribution of the proposed framework is strategic incremental knowledge derivation through topic modelling using LSI and using standard Semantic wikis like Wikidata and knowledge stack repositories like YAGO for entity enrichment in the proposed framework which is quite novel. The presence of a CNN classifier and XGBoost classifier to classify the dataset and the generated dynamic knowledge stack from Web 3.0 is also quite new. The generation of a dynamic knowledge stack from the crux of Web 3.0 is not only novel to the framework but also increases the density of auxiliary knowledge and thereby yields faster convergence to optimality. The presence of Simpson's diversity Index and SOC-PMI as Semantic-similarity measures at different stages in the architectural pipeline helps in Semantic-oriented reasoning and learning , soccer league competition algorithm which is the best in class in metaheuristic which adds to the nobility of the framework.

1.3. Organization

Section 1 depicts the Introduction, Section 2 presents the Related Works, Section 3 outlines the Proposed System Architecture, Section 4 covers the Implementation, Section 5 presents Performance Evaluation and Results, Section 6 covers Discussion, and Section 7 depicts Conclusion.

2. Related Works

Ilyas et al.[1] introduce a novel method combining video scene ontology and CNN for superior video tagging, attaining 99.8% accuracy and a 96.25% F1-score on the UCF-101 dataset, efficiently extracting content from key frames to generate frequent tags and summarize video content. Fernandez et al.[2] put forth ViTS, an automatic Video Tagging System, that enhances video annotations by leveraging Web content and social network comments, maintaining a dynamic knowledge base, indexing videos with multiple labels, and achieving an impressive 80.87% accuracy, surpassing YouTube-8M with publicly available tags and video summaries. Bianco et al.[3] proposed ViTS, utilizes Web content and social network comments to enrich annotations, boasting an 80.87% accuracy and outperforming YouTube-8M with a dynamic knowledge base and multiple labels. Wu et al.[4] present an innovative model for video tagging that includes temporal aspects and personalization, incorporating temporal dependencies, user interactions, and preferences, with extensive analyses confirming superior performance in tag extraction from noisy user-generated comments on large datasets.

Yao et al.[5] explore user search behavior using click-through data, employing co-click statistics and polynomial semantic indexing for economical video annotation, providing a cost-effective alternative with competitive performance in tag assignment, ranking, and enrichment compared to machine learning and data-driven methods. Yang et al.[6] address the challenge of automatic video tagging by introducing the Cross-Media Tag Transfer (CMTT) framework, which leverages well-tagged images to enhance video tagging. It employs a cross-media knowledge transfer approach, optimizing kernel spaces and exploring data structures to improve video classification, outperforming existing algorithms in extensive experiments. Vondrick et al.[7] have proposed a novel active learning framework that introduces efficient video annotation. It selects frames for annotation that can significantly improve object tracking, minimizing user effort. Using a constrained tracker and dynamic programming, the method demonstrates cost-efficient label acquisition on four data sets, including those with key frame annotations from Amazon Mechanical Turk. Larson et al.[8] have given an overview that discusses three tasks from the MediaEval 2010 benchmarking initiative. The tasks involve automatic tagging of Dutch television episodes with subject labels, predicting user-assigned tags for online videos, and automatically assigning geo-coordinates to videos, utilizing a wide range of available information, such as user-generated metadata and audio-visual features.

Pedro et al.[9] analyze redundancy in YouTube videos through content-based techniques, revealing relationships between duplicated content and metadata. It also introduces tag propagation methods to enhance video annotations, validated through experiments and user evaluation. Seirsdorfer et al.[10] discuss how redundancy in YouTube videos can reveal connections between them. It proposes tag propagation methods to enhance video annotations, leading to improved data organization and search, supported by experiments and user evaluations. Wang et al.[11] introduce OMG-SSL, a method for learning-based video annotation that addresses challenges like limited training data and high dimensionality. OMG-SSL utilizes multiple graphs to represent a range of factors in video annotation, optimizing their fusion and weight assignment for enhanced video retrieval performance, as validated in extensive TRECVID benchmark experiments. Rich et al.[12] process recent advances in video annotation tools to enable enhanced teacher self-reflection by facilitating documentation, self-analysis, and tracking development over time, along with linking video to student and teaching evidence. The paper discusses various emerging tools and their potential to revolutionize teacher reflection.

3. Proposed System Architecture

Figure 1 Illustrates the proposed system architecture of the video tag recommendation framework using knowledge-driven hybrid intelligence where the datasets of the video are subjected to extraction of categories and annotations. The categories and annotations are obtained from video labels. Since the categories and annotations are stand-alone entities they need to be

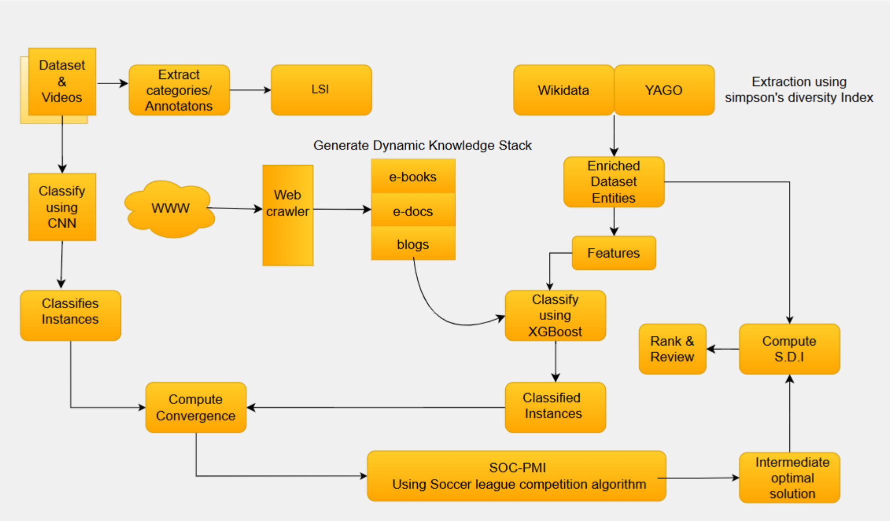

Figure 1: VTKHI Proposed System Architecture Model

subjected to enrichment by augmenting knowledge to it with topic discovery/topic modelling schemes like latent semantic indexing is applied. Latent Semantic indexing is a topic modelling strategy which has the World Wide Web itself as the reference corpora and includes topics to the categories and annotations and contextualizes it. Since the topic for contextualization is quite low in number they need to be further subjected to the addition /augmentation of entities from the World Wide Web such as standard knowledge store repositories like Wikidata and YAGO. Repositories like Wikidata and YAGO house community-sourced and validated cloud source knowledge. The world knowledge is available from which the entities are extracted and augmented to the topics discovered to yield enriched dataset entities.

These enriched dataset entities are subjected to feature extraction onto which Simpson's diversity index is applied. Simpson's diversity index is set to a deviance of 0.15 to select features and these features are sent to the XGboost classifier to classify the dynamically generated knowledge stack sourced from the global Web through crawling. The current structure consists of the e-books, e-docs, and blogs. The Web crawler in the fourth form crawls the entities exactly by using the categories and annotations from the dataset itself. Henceforth, a dynamic knowledge stack is generated owing to more data subjected to classification using the XGboost classifier and the classifier instances are further used. Subsequently, a strong deep learning model namely CNN (convolution neural network) is used to categorize the dataset. CNN is a robust deep learning classifier; its utilization is based on the principle of automatically selecting crafted features and the CNN classifier inputs the actual contents of the videos and the categories of the datasets. Classified instances which come out of a CNN classifier also known as the classified dataset and that of the XGboost classifier which is the classified knowledge stack subjected to computation of conversions using soccer- league competition algorithm. Soccer-league competition algorithm is a meta-heuristic algorithm used as a strategy under the second order co-occurrence point-wise mutual information (SOC-PMI).

SOC-PMI is used as an objective function and sets threshold of 0.75. The threshold is set to 0.75

in SOC-PMI because of the large number of instances that come out of the dynamic knowledge stack and even though SOC-PMI is very strong, to improve the overall relevance of the entities. So, soccer-league competition algorithm is used as a classifier that provides intermediate optimal solution set. This intermediate optimal solution set is subjected to Simpson's diversity index at the said deviance of 0.15 from which the enriched entities dataset is derived to yield the most number of entities and is ranked in the increasing order of the Simpson's diversity index and is sent for review. So, the intermediate optimal solution is present as the first line of tags and the second line of tags which are more specific is the tags that come out of the Simpson's diversity index pipeline. So, the tags that come out through Simpson's diversity index are presented as that of higher priority, which is preceded by the intermediate optimal solution set, this gives a very rich tag set for video tag recommendation using Semantic-oriented knowledge-driven hybrid intelligence.

Simpson's diversity index also known as Dominance Index measures community diversity, although it is mainly used in biodiversity it can also be used to gauge population differences. It refers to 3 sub-divisions

FORMULA:

$$D = \sum \left(\frac{n}{N}\right)^2, \quad D = \sum \frac{n(n-1)}{N(N-1)} \quad \text{where} \quad 0 < D < 1 \tag{1}$$

In equation 1, With a range of D from 0 to 1, the Simpson's index calculates the probability that two people chosen at random from a sample belong to the same species where n indicates the total quantity of a specific species' organisms and N symbolizes the total quantity of all species' organisms. A greater D value indicates less diversity because a D value of 0 indicates infinite diversity and a D value of 1 indicates no diversity. It is quantified how likely it is that two randomly chosen members of a sample are members of distinct species by the Simpson's index of diversity of (1-D). D nevertheless has a value between 0 and 1, despite this a higher D value now denotes more sample diversity. When the Simpson's reciprocal index (1/D) is 1, it indicates that there is only one species present in the community. The community is more diverse the higher the index value.

Within the area of AI and ML, ”XGboost” also referred to as ”Extreme Gradient Boosting” has gained immense recognition. Professionals highly value it due to its remarkable performance and dominance in data science competitions. XGBoost, a member of the ensemble learning family, excels at combining predictions from several weaker models, particularly decision trees, to produce a strong and precise prediction model. Its novel gradient boosting framework, which trains these decision trees repeatedly while optimizing a loss function, enables this. When compared to other algorithms, XGBoost's speed is one of its greatest advantages; it operates incredibly quickly. Additionally, it is very portable and scalable without sacrificing precision or accuracy. The capacity of XGBoost to manage big datasets with ease using parallel processing techniques is what makes it stand out from the competition. Additionally, it uses sophisticated regularization techniques like L1 and L2 regularization, which avoid the overfitting problems that conventional approaches frequently encounter. This algorithm shows itself to be equally effective for both classification and regression tasks, with various applications, spanning a broad spectrum, including finance, healthcare, and natural language processing (NLP), to mention a few.

A Convolutional Neural Network (CNN) classifier, inspired by the occipital cortex in living organisms, is a deep learning architecture designed for image recognition. Revolutionizing computer vision, CNNs automatically learn and distinguish patterns, shapes, and features in images, akin to the human vision system. Comprising layers that perform distinct operations, initial layers focus on textures, detecting contours and shapes. As the network deepens, it captures hierarchical features, recognizing increasingly complex patterns and generating high-level representations. The final layers map these representations to concrete classes, refining through training to minimize prediction differences. CNNs excel in applications beyond image classification, demonstrating impressive performance in object detection, image segmentation, and even natural language processing. Their innate ability to manage spatial hierarchies and extract features positions CNNs as essential tools in modern machine learning, contributing significantly to both research and practical applications.

The soccer league competition algorithm is a new method for optimal solution findings using a metaheuristic technique. It is a bunch of procedures and regulations used to organize and manage a soccer (football) league, regulating how teams play against each other, how points are awarded, and how the overall standings are computed. There are several common algorithms used in soccer league competitions, with the most popular being the round-robin format and variations of it. The Round robin algorithm is as follows:

1. Number of Teams: The primary step is to decide the number of teams taking part in the league. Let us consider there are 'n' teams.
2. Fixtures: In a round-robin competition, each team plays against every other team precisely once. This implies that in each round of matches, there will be n/2 matches. The matches are orchestrated in a way that guarantees each team plays both home and away matches against each opponent.
3. Match Planning: To create a schedule that reduces clashes and guarantees reasonableness, a common approach is to utilize a pattern called a "cyclic algorithm." This algorithm generates the matches by keeping one team settled and rotating the rest of the teams around it. After each turn, the teams are paired up for matches.
4. Scoring and Points: In most soccer leagues, teams are awarded points based on the result of each match. A team receives three points for a win, one point for a draw, and zero points for a loss. The points are used to determine the league standings.
5. League Standings: The league standings are determined by the total points earned by each team. If two or more teams have the same number of points, tiebreakers like goal difference, goals scored, head-to-head results, etc., might be used to determine their ranking.
6. Season Length: The competition generally has multiple rounds, with each round consisting of n/2 matches. The total number of rounds in a season is usually n-1.
7. Promotion and Relegation (In Some Leagues): Depending on the league structure, the teams that perform well might get promoted to a higher division, while the teams at the bottom might face relegation to a lower division.

SOC-PMI is a statistical measure used in natural language processing to explore complex relationships between pairs of items, such as words in a text corpus. Unlike traditional PMI, it considers simultaneous occurrences of pairs, examining the extent to which their co-occurrence exceeds what would be anticipated if the elements were independent. Computationally expressed, this measure provides deeper insights into intricate associations within datasets, helping analysts understand complex interactions between pairs of elements.

FORMULA:

$$PMI(A;B \mid X;Y) = \log_2\left(\frac{P(A,B \mid X,Y)}{P(A \mid X)P(B \mid Y)}\right) \tag{2}$$

4. Implementation

Algorithm 1 Video Tagging Framework using Knowledge-driven Hybrid Intelligence and Quantitative Semantics Driven Reasoning

```
Input: Video Dataset $D$
Output: Ranked Tags for Videos
Begin
    Step 1: Initialization
        $D \leftarrow$ Load Video Dataset
        $R \leftarrow$ Initialize Knowledge Repositories (Wikidata, YAGO)
        $MCNN \leftarrow$ Load CNN Model
        $MXGB \leftarrow$ Load XGBoost Model
    Step 2: Extract Categories and Annotations
        for each video $v \in D$ do
            $C_v \leftarrow$ Extract Categories from $v$
            $A_v \leftarrow$ Extract Annotations from $v$
        end for
    Step 3: Enrich Categories and Annotations
        for each $c \in C_v$ and $a \in A_v$ do
            $T_{c,a} \leftarrow$ Apply LSI to $c$ and $a$
        end for
        $C_v \leftarrow C_v \cup T_{c,a}$
        $A_v \leftarrow A_v \cup T_{c,a}$
        Integrate with $R$
    Step 4: Generate Dynamic Knowledge Stack
        $D_{web} \leftarrow$ Web Crawl for eBooks, Documents
        $K_v \leftarrow$ Assemble Dynamic Knowledge Stack from $D_{web}$
    Step 5: Feature Extraction and Selection
        $F_v \leftarrow$ Extract Features from $C_v$ and $A_v$
        $F'_v \leftarrow$ Apply Simpson's Diversity Index (Threshold: 0.15) to $F_v$
    Step 6: Classification
        Classification$_{XGB} \leftarrow MXGB(K_v)$
        Classification$_{CNN} \leftarrow MCNN(v)$
    **Step 7: Compute Semantic Relatedness**
        **for** each $(f_i, f_j) \in F'_v$ **do**
            SOC-PMI($f_i, f_j$) ← Compute SOC-PMI (Threshold: 0.75)
        **end for**
    **Step 8: Optimize Using Soccer League Competition Algorithm**
        $S_{\text{opt}}$ ← Apply Soccer League Competition Algorithm to SOC-PMI
    **Step 9: Rank and Review Tags**
        $E_v$ ← Apply Simpson's Diversity Index to $S_{\text{opt}}$
        ranked_tags ← Rank Tags in $E_v$
    **Step 10: Output Ranked Tags**
        **for** each $v \in D$ **do**
            Output ranked_tags for $v$
        **end for**
**End**
```

5. Performance Evaluation and Results

These are the Results from the assessment of the suggested (VTKHI) framework video tagging encompassing knowledge-driven hybrid intelligence, which is reviewed using precision, F-measure percentages, accuracy, false discovery rate (FDR), and recall. Potential metrics precision, recall, accuracy, f measure, and FDR are preferred because they quantify the significance of the outcome. FDR quantifies the false positives needed according to the framework and thereby quantifying the error rate of the model, hence move to this preferred as an auxiliary metric.

Table 1
Effectiveness in Performance of VTKHI over other approaches

Model	Average Precision %	Average Recall %	Average Accuracy %	F-Measure %	FDR
DLPT [1]	88.23	91.54	89.88	89.85	0.12
VTPHM [2]	91.23	92.19	91.71	91.70	0.069
MDVC [3]	92.09	93.97	93.03	93.02	0.08
Proposed VTKHI	**96.74**	**97.43**	**97.08**	**97.08**	**0.04**

From Table 1, it is indicated that the proposed VTKHI has an overall average precision of 96.7%, an overall average recall of 97.43%, overall average accuracy of 97.08%, overall F-measure of 97.08%, and an overall FDR of 0.04%. It is indicative from Table 1 that the proposed VTKHI has the highest precision, recall, accuracy, and F-measure percentages and has the lowest value of FDR. To collate the effectiveness of the proposed VTKHI, it serves as a reference point with three distinct models for video tagging frameworks, namely the (DLPT), (VTPHM), (MDVC) respectively. The DLPT yields have yielded 88.23% of precision, 91.54% of recall, 89.88% of accuracy, 89.85% of F-measure and 0.12% of FDR. The VTPHM has furnished a 91.23% of precision, 92.195% of recall, 91.71% of accuracy, 91.70% of F-measure and 0.09% of FDR. The MDVC model has yielded 92.02% of precision, 93.97% of recall, 93.03% of accuracy, 93.02% of F-measure, and 0.08% of FDR.

The rationale behind the proposed VTKHI framework having the highest recall, accuracy, F-measure percentages, precision, and lowest value of FDR, thereby outperforms all the other models because it is based on incrementing knowledge aggregation by harvesting data from categorical data and annotations from the dataset of videos and subsequently enriching through topic modelling using latent Semantic indexing and other knowledge store repositories which is a community contributed and community verified like Wikidata and YAGO. Apart from this, a very strong dynamic knowledge stack is generated through calling the World Wide Web and generating eBooks, documents, and blogs relevant to the categories of the dataset. Apart from this, the proposed VTKHI has a very strong learning infrastructure in terms of convolutional neural networks to categorize the video dataset in the model. Besides this, the soccer league competition algorithm under SOC-PMI, the Simpson's diversity Index and all help in strategic Semantics-oriented reasoning through Semantic similarity measures which are quantitative in nature. So quantitative Semantic reasoning takes place. So, soccer league competition algorithm also helps to generate a better final optimal solution set in the framework. The XGBoost algorithm also provides a very strong learning infrastructure which classifies the dynamic knowledge stack generated by harvesting features through the Simpson's diversity Index from the dataset synthesized of knowledge over YAGO and Wikidata. There is a pipeline going through hybridization of very strong learning infrastructures through CNN and XGBoost and the presence of a strong metaheuristic optimization. For example, soccer league competition algorithm is very strong learning, reasoning infrastructure for SOC-PMI, which helps in quantitative reasoning and Simpson's diversity Index as well as the presence of community contributed knowledge stack repositories like Wikidata, YAGO, a generation of dynamic knowledge stack and topic modelling have a very strong knowledge aggregation infrastructure to the model, making it a very strong hybridized framework for recommending.

The reason why an approach for accurate video tagging using a deep learning model (DLPT) lapses in performance is that although it is a deep learning-based model for precise video tag recommendation that uses the CNN with the video-scene ontology, the scene ontology is quite limited in nature. It does not house or bring about strong auxiliary knowledge density into the model. However, though yields of the knowledge, it quite sparse and the absence of learning and reasoning focused on Semantics only works on the operational dataset, which means factual reasoning is limited in the model, and DLPT model Does not meet expectations when evaluated against the suggested approach.

The reason why the Video Tagging via Probabilistic Hybrid Modelling (VTPHM) also falls short in terms of performance expected when juxtaposed to the proposed framework is that although the VTPHM model is a video tagging through probabilistic hybrid modelling, it uses a hierarchical structure based multi-concept model with regression analysis. The classification model uses both multi-model and temporal properties. However, the video concept dependencies and temporal dynamics must be scaled and correlated, which is not done. Moreover, this framework video tagging with probabilistic hybrid modelling does not have a very strong Semantics-oriented learning and reasoning. The optimization strategy is absent in the model and henceforth it lacks drastically over the proposed framework.

Finally, the third model, the Multi-Modal Dense Video Captioning (MDVC) also fails to meet expectations when evaluated against the proposed framework because although it is a multimodal, dense video captioning framework which uses multimodal strategy of images, text, and speech, also has a transformer architecture. Semantics-orientation is complete, although it uses multimedia and multimodal dense video captioning, The quantity of supplementary information added into the model is quite dubious. Redundant and cognitive knowledge synthesis from the actual Web 3.0 through cognitive and mental states or human conceptual understanding is absent. Non-synonymous learning infrastructure also can be improved, which makes it lag and Semantic-oriented reasoning is absent. Henceforth, the model is not as beneficial as the proposed model.

Table 2
Precision Percentage VS Number of Recommendations

No. of recommendations	DLPT	VTPHM	MDVC	Proposed VTKHI
10	90.15	93.74	94.07	98.09
20	89.07	92.09	93.45	97.09
30	88.45	91.09	92.09	96.72
40	87.09	90.27	91.09	95.12
50	86.55	89.09	90.42	94.07

Table 2 depicts the Precision vs. Number of Recommendations distributions curve as a main graph for the proposed VTKHI and the distinct models that are DLPT, VTPHM, and MDVC respectively. The DLPT occupies the lowest in the hierarchy and VTPHM occupies the last but one in the hierarchy because of their disadvantages. MDVC occupies second from the top because of its increased computational complexity due to the integration of diverse modalities, potentially leading to higher resource requirements and processing time. The reason why the proposed framework shows better results is due to all the positives of proposed VTKHI model, over the baseline models and since the proposed VTKHI has a hybrid Semantic and intelligent framework which incrementally aggregates knowledge and has a very strong learning infrastructure in terms of CNN, the XGBoost classifier classifying the dataset and dynamically generated knowledge stack. Since it has a very strong knowledge infrastructure through dynamic knowledge stack generation. Wikidata and YAGO are LSI-based topic models and knowledge-based repositories. Since the presence of SOC-PMI and Simpson's diversity Index of quantitative Semantics-oriented similarity and soccer league competition algorithm to serve as a potential metaheuristic optimization model that helps this framework from all the baseline models, which serve as a best-in-class model for recommending video tags to videos.

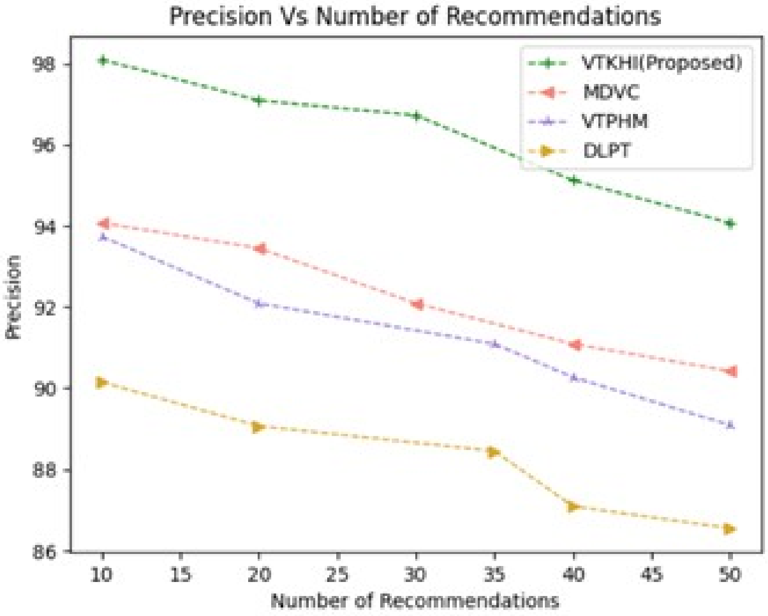

Figure 2: Precision Vs Number of Recommendations Graph

6. Discussion

The suggested VTKHI framework naturally accommodates multilingualism, which is essential for promoting inclusivity and significance in the Web 3.0 paradigm. Utilizing Semantic wikis such as Wikidata and YAGO, which contain extensive multilingual data contributed by the community, the framework facilitates strong enhancement of video annotations in multiple languages. Employing quantitative Semantic reasoning metrics like SOC-PMI and Simpson's Diversity Index guarantees language-neutral processing, enabling the framework to effectively analyze and contextualize datasets irrespective of linguistic variations. Moreover, the CNN and XGBoost classifiers emphasize features that go beyond language differences, enabling the system to adjust to various linguistic settings while preserving its high accuracy and efficiency.

The multilingual capabilities of this framework have the potential to significantly impact international markets in the future by enhancing accessibility and video tagging for diverse audiences. Enhancing cross-lingual tagging can help content creators reach multilingual audiences and improve content discoverability. Combining this framework with sophisticated multilingual natural language processing models can improve tag accuracy and cultural significance as AI-based semantic comprehension advances, supporting Web 3.0's inclusive and interconnected vision.

7. Conclusion

This paper proposes a video tagging recommendation framework that encompasses hybrid intelligence and auxiliary knowledge addition in which topic modelling like latent Semantic indexing (LSI) helps in the contextualization of the initially discovered or initially expected terms from the dataset. The Wikidata and YAGO knowledge store repositories help in contributing to the amount of lateral auxiliary knowledge that is added to the framework from community-contributed or community-verified resources to enrich the dataset which indirectly increases the cognition of the proposed framework. CNN is used to classify the dataset of videos and XGBoost to classify the generated dynamic knowledge stack which helps in transforming the dataset and the dynamic knowledge stack to be more atomic and accommodative to the localized framework. Convergence computation using soccer league competition as a metaheuristic with SOC-PMI as a criteria function not only serves as a very strong quantitative Semantic reasoning approach but also is the best-in-class intermediate optimal solution set. Simpson's diversity Index computation with an empirically decided step deviance measure also promotes the addition and regulation of auxiliary knowledge and facilitates Semantics-oriented reasoning through quantitative Semantic-similarity measure. An overall precision of 96.74

References

[1] S. Ilyas, H. U. Rehman, A deep learning based approach for precise video tagging, in: 2019 15th International Conference on Emerging Technologies (ICET), IEEE, 2019.

[2] D. Fernández, et al., Vits: video tagging system from massive web multimedia collections, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017.

[3] S. Bianco, et al., An interactive tool for manual, semi-automatic and automatic video annotation, Computer Vision and Image Understanding 131 (2015) 88–99.

[4] B. Wu, et al., Crowdsourced time-sync video tagging using temporal and personalized topic modelling, in: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014.

[5] T. Yao, et al., Annotation for free: Video tagging by mining user search behaviour, in: Proceedings of the 21st ACM international conference on Multimedia, 2013.

[6] Y. Yang, Y. Yang, H. T. Shen, Effective transfer tagging from image to video, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 9 (2013) 1–20.

[7] C. Vondrick, D. Ramanan, Video annotation and tracking with active learning, in: Advances in Neural Information Processing Systems 24, 2011.

[8] M. Larson, et al., Automatic tagging and geotagging in video collections and communities, in: Proceedings of the 1st ACM international conference on multimedia retrieval, 2011.

[9] J. San Pedro, S. Siersdorfer, M. Sanderson, Content redundancy in youtube and its application to video tagging, ACM Transactions on Information Systems (TOIS) 29 (2011) 1–31.

[10] S. Siersdorfer, J. San Pedro, M. Sanderson, Automatic video tagging using content redun-

dancy, in: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009.

[11] M. Wang, et al., Unified video annotation via multigraph learning, IEEE Transactions on Circuits and Systems for Video Technology 19 (2009) 733–746.

[12] P. J. Rich, M. Hannafin, Video annotation tools: Technologies to scaffold, structure, and transform teacher reflection, Journal of Teacher Education 60 (2009) 52–66.