جامعة سيدي محمد بن عبد الله بفاس
UNIVERSITE SIDI MOHAMED BEN ABDELLAH DE FES
كلية العلوم ظهر المهراز فاس
FACULTE DES SCIENCES DHAR EL MAHRAZ FES

CENTRE D'ETUDES DOCTORALES
«SCIENCES ET TECHNIQUES ET SCIENCES MÉDICALES »
مركز الدكتوراه « العلوم والتقنيات والعلوم الطبية»

AVIS DE SOUTENANCE DE THESE

Le Doyen de la Faculté des Sciences Dhar El Mahraz –Fès – annonce que

Mr **KHELDOUNI Mohammed Amine**
Soutiendra : **le Vendredi 24/07/2026** à **09H30**
Lieu : ***FSDM – Centre Visioconférence***

Une thèse intitulée :

From Collaborative Signals to Sensory Content: Multimodal Approaches for Enhanced and Generalizable Recommender Systems

En vue d'obtenir le ***Doctorat***

FD : **Sciences et Technologies de l'Information et de la Communication**
Spécialité : **Informatique &Intelligence Artificielle**

Devant le jury composé comme suit :

Nom et prénom	Etablissement	Grade	Qualité
Pr. NFAOUI El Habib	Faculté des Sciences Dhar El Mahraz, Fès	PES	Président
Pr. ALIPPI Cesare	Politecnico di Milano, Italie	PES	Rapporteur
Pr. OUHBI Brahim	Ecole Nationale Supérieure des Arts et Métiers, Meknès	PES	Rapporteur
Pr. AMAKDOUF Hicham	Faculté des Sciences Dhar El Mahraz, Fès	MCH	Rapporteur
Pr. TAIRI Hamid	Faculté des Sciences Dhar El Mahraz, Fès	PES	Examinateur
Pr. ZAHI Azeddine	Faculté des Sciences et Techniques, Fès	PES	Examinateur
Pr. BOUMHIDI Jaouad	Faculté des Sciences Dhar El Mahraz, Fès	PES	Directeur de thèse

Résumé :

Les systèmes de recommandation sont des composants essentiels des plateformes numériques modernes. Cependant, le paradigme dominant du filtrage collaboratif (CF) traite les articles comme des identifiants opaques, ignorant le contenu sensoriel riche à travers lequel les utilisateurs forment réellement leurs préférences. Cette thèse propose et valide un paradigme multisensoriel de la recommandation qui intègre systématiquement des modalités de contenu — texte, image, audio et vidéo — dans des architectures de filtrage collaboratif neuronal, alignant les représentations des articles avec les canaux perceptifs de l'expérience humaine.

La thèse s'articule autour de trois contributions complémentaires qui élargissent progressivement l'espace sensoriel exploité par le système de recommandation.

1. V-BERT4Rec démontre que les caractéristiques visuelles extraites des images d'articles améliorent la recommandation séquentielle auto-attentive. En injectant des plongements d'images issus d'un Vision Transformer gelé dans l'architecture BERT4Rec via une fusion additive des embeddings, V-BERT4Rec surpasse de manière cohérente les baselines basées uniquement sur les identifiants sur MovieLens 100K et 1M, avec les gains les plus importants sur le jeu de données le plus sparse.

2. RecCLIP s'attaque au problème du démarrage à froid inhérent aux contenus générés par intelligence artificielle, pour lesquels aucun historique d'interaction n'existe. Des embeddings d'utilisateurs apprenables sont combinés avec les encodeurs gelés CLIP-L/14 (texte et image) via une attention croisée de type Transformer et entraînés avec une perte d'entropie croisée binaire. RecCLIP surpasse PickScore et DPO-SDXL sur le benchmark Pick-a-Pic pour la prédiction personnalisée de préférences.

3. MM-NCF propose une évaluation systématique de quatre modalités de contenu dans un cadre de filtrage collaboratif unifié. Les caractéristiques textuelles (Sentence-BERT), visuelles (CLIP), audio (Whisper) et vidéo (CLIP sur trames échantillonnées) sont intégrées dans le filtrage collaboratif neuronal via des projections à portes de qualité, une fusion hiérarchique à deux niveaux (groupes sémantiques vs. perceptuels) et un apprentissage par curriculum adaptatif au jeu de données. L'ablation systématique sur MovieLens 100K et 1M révèle que (i) les quatre modalités améliorent le classement sur les jeux de données peu denses (sparses), (ii) les modalités perceptuelles (audio, vidéo) restent bénéfiques même sur les jeux denses, (iii) le taux de dropout de l'embedding d'article est un paramètre de contrôle critique dont l'optimum diminue avec la densité des interactions, et (iv) l'augmentation multimodale produit des effets hétérogènes mais majoritairement positifs selon les groupes démographiques et les catégories de films. La comparaison avec les méthodes de l'état de l'art (VBPR, MMGCN, LATTICE, FREEDOM, BM3) confirme des performances compétitives ou supérieures, tout en étant le seul système à intégrer les modalités audio et vidéo, à prédire des notes explicites et à fournir une analyse par sous-groupes.

En synthétisant les trois contributions, cette thèse établit quatre résultats clés :
(1) les modalités de contenu encodent des informations complémentaires au filtrage collaboratif (CF), les modalités perceptuelles capturent des qualités dynamiques peu susceptibles d'émerger des schémas de co-occurrence ;

(2) le bénéfice des caractéristiques de contenu dépend de la densité des données, saturant pour les modalités sémantiques dans les jeux de données denses ;
(3) l'équilibre CF–contenu est explicitement contrôlable via le dropout de l'embedding d'article ; et
(4) l'évaluation par sous-groupes révèle que les caractéristiques de contenu bénéficient de manière disproportionnée aux genres visuellement distinctifs et aux segments d'utilisateurs sous-représentés.

Les perspectives incluent la validation inter-domaines, les schémas de dropout adaptatifs, des modalités sensorielles plus riches (olfactives, haptiques) et des modèles de fondation multimodaux de bout en bout pour la recommandation.

Mots clés :
systèmes de recommandation, apprentissage multimodal, filtrage collaboratif, recommandation séquentielle, caractéristiques visuelles, caractéristiques audio, caractéristiques vidéo, CLIP, fusion à portes de qualité, recommandation à froid.

FROM COLLABORATIVE SIGNALS TO SENSORY CONTENT: MULTIMODAL APPROACHES FOR ENHANCED AND GENERALIZABLE RECOMMENDER SYSTEMS

Abstract :

Collaborative filtering (CF), the dominant paradigm in modern recommender systems, treats items as opaque identifiers and discards the content through which users form preferences. This thesis investigates a multimodal approach to recommendation that integrates four content modalities — text, image, audio, and video — into neural collaborative filtering, grounding item representations in perceptual signals rather than interaction patterns alone.

Three contributions progressively expand the modality space.

1. V-BERT4Rec injects image embeddings from a frozen Vision Transformer (ViT-Base/16) into the BERT4Rec sequential encoder via additive embedding fusion. The model consistently outperforms ID-only baselines on MovieLens 100K and 1M, with the largest gains on the sparser dataset where collaborative signal is the weakest.

2. RecCLIP targets the cold-start setting of AI-generated content, where no interaction history exists. Learnable per-user embeddings are combined with frozen CLIP-L/14 text and image encoders via Transformer-based cross-attention and trained with binary cross-entropy loss. RecCLIP outperforms PickScore and DPO-SDXL on the Pick-a-Pic benchmark for personalised preference prediction.

3. MM-NCF integrates text (Sentence-BERT), image (CLIP), audio (Whisper), and video (frame-sampled CLIP) into Neural Collaborative Filtering through quality-gated projections and hierarchical two-level fusion. Ablation on MovieLens 100K and 1M shows that (i) all four modalities improve ranking on sparse data, (ii) audio and video remain beneficial even on denser data where text and image saturate, (iii) item embedding dropout controls the CF–content balance with an optimum that decreases as interaction density grows, and (iv) subgroup analysis by gender, age, and movie genre reveals heterogeneous but broadly positive effects. Benchmarking against VBPR, MMGCN, LATTICE, FREEDOM, and BM3 confirms competitive or superior performance, while MM-NCF uniquely supports audio and video modalities.

Four cross-cutting findings emerge:
(1) content modalities encode information that CF cannot recover from co-occurrence patterns alone, particularly for dynamic qualities captured by audio and video;
(2) the marginal value of content features decreases with interaction density, saturating first for semantic modalities;
(3) item embedding dropout provides an explicit, single-parameter mechanism for tuning the CF–content balance; and
(4) the benefits of multimodal augmentation are unevenly distributed across genres and demographic groups, favouring visually distinctive categories and users with shorter histories.

Key Words :
recommender systems, multimodal learning, collaborative filtering, sequential recommendation, visual features, audio features, video features, CLIP, quality-gated fusion, cold-start recommendation.