


République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université AMO de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Projet System De Recherche D'Information

Spécialité : *Master 1 ISIL*

Thème 03

Multimedia Information retrieval

Réalisé par :

- Mazouz Yaniz
- Draifi Abdelhak
- Rahal Lyes

2024/2025

Table des matières

]

Introduction

Les ordinateurs, longtemps utilisés pour traiter des informations textuelles et numériques, jouent désormais un rôle central dans l'accès et la gestion d'informations multimédia. En effet, de nombreux domaines professionnels nécessitent des contenus non textuels pour répondre à des besoins spécifiques. Par exemple, les médecins consultent des radiographies, les architectes utilisent des plans de bâtiments, et les agents immobiliers montrent des photographies de propriétés.

Dans ces domaines, l'information visuelle est souvent aussi importante, voire plus, que le texte. Le besoin de recherche d'informations multimédia devient ainsi essentiel, car il permet d'accéder aux documents visuels ou audio nécessaires à ces professionnels. Il est difficile d'imaginer qu'une entreprise de construction reçoive un plan de bâtiment uniquement sous forme textuelle, ou qu'un journal soit consulté sans la disposition graphique de ses pages. La recherche d'informations multimédia devient aussi essentielle dans des contextes où les documents sont principalement textuels mais nécessitent des annotations ou illustrations visuelles, comme les formulaires d'assurance comportant des commentaires en marge.

Les progrès récents en matière de stockage et d'affichage numérique facilitent l'intégration de contenus multimédia dans les documents informatiques, et le traitement des images, vidéos, et sons devient plus accessible. La recherche d'informations multimédia permet aux utilisateurs de créer et de naviguer dans des bibliothèques de documents enrichis, où textes et médias se complètent pour offrir des perspectives variées et améliorer la qualité des informations accessibles.

]

Chapitre 1

Concepts fondamentaux de MIR

Les concepts fondamentaux de la Récupération d'Information basée sur la Recherche de Médias (MIR – Multimedia Information Retrieval) englobent plusieurs notions clés qui permettent de rechercher et de récupérer des informations multimédia (textes, images, audio, vidéo, etc.) de manière efficace. Voici un aperçu des concepts essentiels dans ce domaine :

1.2 Indexation multimodale

L'indexation est le processus d'attribution de métadonnées aux objets multimédia afin qu'ils puissent être rapidement récupérés lors d'une requête. L'indexation multimodale fait référence à l'organisation de différents types de médias (textes, images, vidéos, audio) de manière à permettre une recherche efficace sur l'ensemble de ces données.

1.2.1 Indexation textuelle

L'indexation textuelle consiste à associer des mots-clés ou des métadonnées à des documents textuels pour en faciliter la recherche et l'organisation. Elle repose sur l'identification des termes significatifs ou expressions clés qui représentent le contenu, tels que énergie solaire ou transition énergétique pour un article sur les énergies renouvelables. Les métadonnées peuvent inclure des informations comme l'auteur, la date de création, les catégories thématiques ou un résumé descriptif. L'indexation peut être effectuée manuellement par des experts ou automatiquement à l'aide d'algorithmes utilisant des techniques de traitement automatique du langage naturel (TALN). Ce processus améliore

considérablement l'accès à l'information, permet de retrouver rapidement des documents pertinents dans de grandes bases de données et optimise les moteurs de recherche. Par exemple, pour un document intitulé "L'intelligence artificielle dans la médecine moderne", les mots-clés pourraient inclure IA, médecine, ou diagnostic assisté par ordinateur.

1.2.2 Indexation visuelle

L'indexation visuelle consiste à extraire et analyser les caractéristiques visuelles des images et vidéos pour les organiser, les classifier et faciliter leur recherche. Ces caractéristiques peuvent inclure des informations telles que les couleurs dominantes, les textures, les formes ou encore les objets reconnus au sein des images. Par exemple, une image contenant un coucher de soleil pourra être associée à des teintes chaudes (orange, rouge), des textures de ciel et des formes naturelles comme des montagnes. L'indexation visuelle repose souvent sur des algorithmes d'apprentissage automatique ou de vision par ordinateur qui permettent d'identifier et d'étiqueter automatiquement les contenus visuels. Ce processus est largement utilisé dans des domaines tels que les bibliothèques d'images, les moteurs de recherche visuelle, ou encore les systèmes de surveillance, permettant une navigation intuitive et efficace parmi des collections visuelles volumineuses.

1.2.3 Indexation audio

L'indexation audio implique l'identification et l'extraction de caractéristiques spécifiques des fichiers audio pour en faciliter l'organisation et la recherche. Parmi ces caractéristiques, on trouve les fréquences, qui permettent d'identifier les tonalités ou les gammes, et les rythmes, qui aident à décrire le tempo et la structure temporelle du son. D'autres éléments comme les timbres (ou les qualités sonores) et les motifs musicaux peuvent également être analysés pour mieux comprendre le contenu audio. Cette indexation est souvent effectuée à l'aide d'algorithmes de traitement du signal et d'apprentissage automatique qui reconnaissent des schémas audio distinctifs, facilitant ainsi la catégorisation de fichiers musicaux, les recherches par contenu, ou l'identification d'échantillons sonores dans de grandes bases de données.

1.3 Recherche d'informations multimédia

Dans MIR, les informations sont représentées sous des formats spécifiques pour faciliter leur récupération. Cela inclut la représentation de données multimédia sous forme de vecteurs ou de signatures, souvent avec des caractéristiques extraites des éléments multimédias.

1.3.1 Modèle vectoriel

Le modèle vectoriel pour la recherche d'informations multimédia représente les documents et les requêtes sous forme de vecteurs dans un espace multidimensionnel. Chaque dimension correspond à une caractéristique (termes, couleurs, fréquences, etc.), et la pertinence est calculée en mesurant la similarité, souvent via le cosinus de l'angle entre les vecteurs. Ce modèle est largement utilisé pour classer et retrouver des contenus en fonction de leur similarité avec la requête

1.3.2 Modèle probabiliste

Il repose sur l'idée d'estimer la probabilité qu'un document soit pertinent pour une requête spécifique. Chaque document est représenté par un ensemble de caractéristiques (termes textuels, éléments visuels ou audio), et un score de pertinence est calculé en fonction de ces caractéristiques et de leur importance pour l'utilisateur. Les documents sont ensuite classés en ordre décroissant de probabilité pour retourner les résultats les plus pertinents en priorité. Ce modèle s'adapte aux incertitudes inhérentes à la recherche et est souvent amélioré par des techniques comme le feedback de pertinence ou l'apprentissage automatique. Il est particulièrement efficace pour intégrer différents types de données multimodales.

1.4 Requête multimodale

Les requêtes multimodales dans la Recherche d'Informations Multimodales (MIR) sont des requêtes où plusieurs types de médias ou modalités sont utilisés ensemble pour améliorer la précision des résultats de recherche. Ces modalités peuvent inclure du texte, des images, des vidéos, des éléments audio, des gestes, ou d'autres types de données. Ce

type de recherche est particulièrement utile dans des contextes où une seule modalité, comme le texte ou l'image, ne permet pas de capturer suffisamment d'informations pour retourner des résultats pertinents

1.4.1 Processus de gestion des requêtes multimodales

1 -Fusion des modalités : Lorsqu'une requête multimodale est lancée, les différents types de médias sont traités et intégrés de manière cohérente. Par exemple, dans le cas d'une requête comprenant un texte et une image, un système peut utiliser des techniques de fusion de caractéristiques, où des vecteurs de caractéristiques représentant le texte et l'image sont combinés. Ensuite, la similarité entre la requête et les documents dans la base de données est calculée sur la base de cette fusion. De même, pour des requêtes qui incluent à la fois des vidéos et des éléments audio, des modèles de traitement multimodal peuvent être utilisés pour extraire des caractéristiques pertinentes de chaque modalité (par exemple, des objets détectés dans la vidéo et des fréquences spécifiques dans l'audio) et pour les intégrer dans un espace de recherche commun.

2 -Alignement sémantique :L'un des défis majeurs des requêtes multimodales est d'assurer un alignement sémantique entre les différentes modalités. Par exemple, une image peut contenir des objets ou des scènes qui ne sont pas directement exprimés par des mots dans une requête textuelle, et vice versa. Des techniques avancées comme les réseaux neuronaux multimodaux sont utilisées pour apprendre les relations entre les différentes modalités (par exemple, associant un objet visuel à des mots-clés dans une requête textuelle). Le but est de s'assurer que les informations provenant de différentes modalités sont correctement interprétées pour que le système puisse comprendre ce que l'utilisateur recherche, même si les termes textuels ne correspondent pas exactement aux éléments visuels.

3 -Recherche basée sur des requêtes combinées : Par exemple, un utilisateur peut effectuer une recherche multimodale en combinant un texte descriptif avec une image ou une vidéo. Un exemple concret serait une recherche dans une base de données d'images où l'utilisateur soumet un texte comme « montagnes au coucher du soleil » et une image représentant un paysage montagneux. Le système doit être capable d'analyser à la fois le texte pour en extraire les mots-clés (comme "montagnes" et "coucher du soleil") et l'image pour en détecter les éléments visuels pertinents (comme les montagnes et les couleurs

chaudes). Ensuite, il associera ces informations pour retourner des résultats correspondant aux deux modalités.

4 -Traitement des données audio et vidéo : Dans des cas plus complexes, comme la recherche multimodale dans des vidéos, la requête peut inclure à la fois du contenu visuel et audio. Par exemple, un utilisateur pourrait soumettre un extrait audio d'une chanson et une image d'un artiste pour retrouver des vidéos associées à cette chanson ou à cet artiste. Le système devra traiter le son pour identifier des motifs ou des fréquences musicales, et analyser l'image pour en extraire des éléments visuels tels que des visages ou des logos, avant de faire correspondre les deux types de données à des vidéos existantes dans la base de données.

1.5 Mesure de la performance

La mesure de la performance dans les systèmes de recherche d'informations multimédia (MIR) est cruciale pour évaluer l'efficacité et la pertinence des résultats retournés. Ces systèmes cherchent à fournir des résultats multimodaux (texte, image, audio, vidéo) en réponse à une requête donnée

1.5.1 Précision

La précision est une mesure qui indique la proportion de documents ou objets pertinents parmi les résultats retournés par le système de recherche. Elle est exprimée par la formule suivante :

$$\text{Précision} = \frac{\text{Nombre de documents pertinents récupérés}}{\text{Nombre total de documents récupérés}}$$

En d'autres termes, elle mesure combien de résultats récupérés par le système sont réellement pertinents par rapport à l'ensemble des résultats retournés. Une précision élevée signifie que le système fournit principalement des résultats pertinents, mais cela ne garantit pas nécessairement que tous les documents pertinents ont été trouvés.

1.5.2 Rappel

Le rappel, également connu sous le nom de sensibilité ou de couverture, est une mesure qui évalue la capacité du système à retrouver tous les documents pertinents disponibles dans la base de données. Il est défini comme la proportion de documents pertinents récupérés parmi tous les documents pertinents existants. La formule du rappel est la suivante :

$$\text{Rappel} = \frac{\text{Nombre de document pertinents récupérés}}{\text{Nombre total de documents pertinents}}$$

Un **rappel élevé** signifie que le système est capable de retrouver une grande partie des documents pertinents dans la base de données. Cependant, cela peut entraîner des résultats moins précis si le système retourne également des documents non pertinents.

1.5.3 F-mesure

La F-mesure (ou F-score) est une mesure combinée qui cherche à équilibrer la précision et le rappel. Elle est particulièrement utile lorsque ces deux critères sont en conflit. La F-mesure est la moyenne harmonique entre la précision et le rappel, ce qui lui permet de donner une évaluation unique tout en tenant compte des deux dimensions. La formule de la F-mesure est la suivante :

$$\text{F-mesure} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

La F-mesure est importante car elle fournit une vue d'ensemble de la performance du système, en s'assurant qu'un équilibre est trouvé entre récupérer des résultats pertinents (précision) et retrouver la majorité des résultats pertinents (rappel). Une **F-mesure élevée** indique un bon compromis entre les deux, ce qui est essentiel dans un système MIR où les utilisateurs veulent à la fois des résultats pertinents et une couverture maximale des éléments pertinents.

]

Chapitre 2

techniques de récupération basée sur le contenu (CBR)

Les techniques de récupération basées sur le contenu permettent d'identifier, de rechercher et de classer des éléments en analysant leurs caractéristiques intrinsèques. Contrairement aux approches traditionnelles qui se fondent sur des mots-clés ou des métadonnées (comme les tags ou descriptions), ces méthodes analysent directement le contenu des éléments, qu'il s'agisse de données visuelles, auditives ou textuelles.

Les **principales étapes** du processus incluent :

1. **Extraction de caractéristiques** : Conversion des éléments multimédias en des représentations numériques compactes appelées descripteurs (par exemple, histogrammes de couleurs, spectrogrammes pour l'audio, etc.).
2. **Indexation** : Organisation des descripteurs dans une structure de données adaptée (par exemple, arbre de recherche).
3. **Recherche et comparaison** : Comparaison des descripteurs d'une requête avec ceux des éléments stockés en utilisant des mesures de similarité.

2.2 Applications pratiques

2.2.1 Recherche d'images

Techniques utilisées :

— Histogrammes de couleurs : Mesures statistiques des couleurs dominantes d'une image.
— Textures : Identification des motifs répétitifs dans une image (par exemple, lignes, points).
— Formes : Analyse des contours et des structures géométriques.
— Deep Learning : Utilisation de réseaux de neurones convolutionnels (CNN) pour extraire des caractéristiques complexes.

Exemple : Recherche d'images similaires à partir d'un croquis ou d'une photo.

2.2.2 Recherche audio

Techniques utilisées :

— Spectrogrammes : Représentation visuelle des fréquences dans un signal audio.
— Fingerprints audio : Création d'une empreinte unique basée sur des fréquences et des rythmes.
— Apprentissage automatique : Modèles d'apprentissage supervisé pour identifier des genres ou des instruments.

Exemple : Retrouver une chanson en fredonnant une mélodie (Shazam).

2.2.3 Recherche vidéo

Techniques utilisées :

— Analyse de frames : Extraction de caractéristiques à partir d'images individuelles dans une vidéo.
— Reconnaissance de scènes : Identification de contextes ou objets spécifiques dans une séquence.
— Indexation spatio-temporelle : Analyse des mouvements et changements dans le temps.

Exemple : Recherche de vidéos sportives contenant un type de but spécifique.

2.2.4 Recherche textuelle

Techniques utilisées :

— TF-IDF (Term Frequency-Inverse Document Frequency) : Évaluation de l'importance des mots dans un document.
— Word Embeddings : Représentation vectorielle des mots (ex. Word2Vec, GloVe).
— Recherche sémantique : Utilisation de modèles NLP pour comprendre le contexte.

Exemple : Rechercher des articles similaires à un texte donné.

2.3 Méthodologies détaillées

2.3.1 Extraction de caractéristiques

Chaque type de contenu possède des descripteurs spécifiques :

— **Images** : Histogrammes, gradients orientés (HOG), descripteurs locaux comme SIFT ou SURF.
— **Audio** : Coefficients cepstraux (MFCC), analyses FFT (Fast Fourier Transform).
— **Vidéo** : Fusion des caractéristiques spatiales (images) et temporelles (mouvements).
— **Texte** : Analyse syntaxique, fréquence des mots, bigrammes/trigrammes.

2.3.2 Mesures de similarité

Les mesures varient en fonction du type de données :

— Distance Euclidienne : Mesure classique entre vecteurs.
— Cosine Similarity : Pour comparer des documents textuels.
— Intersection d'histogrammes : Adaptée aux données d'image.
— Distance de Levenshtein : Pour évaluer les différences entre chaînes de caractères.

2.3.3 Apprentissage automatique

Utilisation de techniques modernes pour améliorer la pertinence :

— Clustering : Organisation non supervisée des données (ex. K-Means).
— Classification supervisée : Modèles comme les SVM ou les réseaux neuronaux pour catégoriser les données.
— Deep Learning : Approches complexes basées sur des réseaux de neurones profonds (CNN pour les images, RNN/LSTM pour les séquences).

2.4 Avantages des CBR

1. Recherche basée sur le contenu réel plutôt que des étiquettes parfois erronées.
2. Permet la recherche d'éléments visuels, sonores ou textuels avec une précision accrue.
3. Exploitation des capacités d'intelligence artificielle pour s'adapter à des bases de données massives.

2.5 Défis et limites

1. Complexité calculatoire : Traitement et comparaison des descripteurs peuvent être coûteux pour de larges bases de données.
2. Qualité des caractéristiques : Une mauvaise extraction peut nuire à la pertinence des résultats.
3. Subjectivité : Les descripteurs ne capturent pas toujours la perception humaine (exemple : beauté subjective dans une image).

]

Chapitre 3

Apprentissage Automatique et Approches Profondes pour le MIR

Les techniques modernes d'Apprentissage Automatique et d'Apprentissage Profond jouent un rôle clé dans la récupération d'informations musicales (Music Information Retrieval - MIR). Elles permettent d'analyser et de traiter automatiquement des données musicales pour des tâches variées.

3.1 Apprentissage Automatique (Machine Learning)

Les approches traditionnelles se basent sur l'extraction de caractéristiques musicales spécifiques avant d'utiliser des modèles pour les analyser.

Caractéristiques couramment utilisées

— **MFCC (Coefficients cepstraux)** : Détectent les propriétés acoustiques d'un son.

— **Descripteurs rythmiques et harmoniques** : Analyser les motifs de rythme et de tonalité.

Modèles classiques

— **KNN et SVM** : Pour classer les genres musicaux ou reconnaître les instruments.

— **Hidden Markov Models (HMM)** : Utilisés pour la transcription musicale et l'analyse séquentielle.

Applications

— Classification des genres musicaux.

— Reconnaissance des instruments.

— Analyse des émotions musicales.

3.2 Apprentissage Profond (Deep Learning)

Les réseaux neuronaux profonds permettent d'analyser directement les signaux audio pour capturer des caractéristiques complexes sans extraction préalable.

Techniques principales

— **CNN (Convolutional Neural Networks)** : Pour analyser les spectrogrammes et détecter des motifs acoustiques.

— **RNN et LSTM (Long Short-Term Memory)** : Pour modéliser les relations temporelles dans la musique.

— **GANs (Generative Adversarial Networks)** : Pour générer de nouvelles compositions musicales.

Applications

— Recherche de morceaux similaires.

— Annotation automatique des émotions.

— Création musicale et transcription.

3.3 Comparaison

Aspect	**Apprentissage Automatique**	**Apprentissage Profond**
Extraction manuelle	Nécessaire	Automatique
Performances	Moyenne	Élevée sur de grands ensembles
Complexité	Simple	Nécessite plus de ressources

]

Chapitre 4

Fusion de Modèles Multimodaux pour l'Amélioration de la Recherche

4.1 Introduction à la fusion multimodale

La récupération multimodale intègre des informations provenant de différents types de données (texte, images, audio, vidéo) pour améliorer les performances des systèmes de recherche. L'objectif est de surmonter les limitations des approches unimodales en exploitant les complémentarités entre les modalités.

4.2 Techniques de fusion

4.2.1 Fusion précoce (Early Fusion)

- Combine les caractéristiques extraites de différentes modalités avant le traitement.
- Par exemple, concaténer des vecteurs représentant du texte et des images dans un espace commun.
- **Avantage** : garantit une interaction riche entre les modalités dès le début.
- **Limite** : les modèles doivent gérer des espaces de données complexes.

4.2.2 Fusion tardive (Late Fusion)

- Combine les résultats des modèles unimodaux après leur traitement séparé.
- Exemple : Combinaison des scores de similarité textuelle et visuelle.

— **Avantage** : facile à mettre en œuvre.
— **Limite** : risque de perte d'information intermodale.

4.2.3 Fusion hybride

— Mixte entre les approches précoces et tardives.
— Idéal pour capturer à la fois les interactions profondes et les relations indépendantes entre les modalités.

4.3 Importance des modèles d'apprentissage profond

Les réseaux neuronaux multimodaux, comme le modèle CLIP (Contrastive Language-Image Pretraining), ont montré leur capacité à aligner efficacement des données hétérogènes dans un espace sémantique partagé.

4.4 Défis et perspectives

— **Alignement sémantique des données** : garantir que les modalités partagent une représentation cohérente.
— **Calcul intensif** requis par les modèles multimodaux.
— **Perspectives** : avancées en transfert d'apprentissage et en modèles génératifs comme les transformers multimodaux.

4.5 Évaluation des Systèmes de Récupération Multimédia

4.5.1 Critères d'évaluation

Un système de récupération multimédia est évalué selon sa capacité à fournir des résultats pertinents et rapides. Les critères incluent :

1. **Pertinence** : Mesure si les résultats correspondent aux intentions de l'utilisateur.
2. **Précision et rappel (Precision & Recall)** :
   — **Précision** : proportion de résultats pertinents parmi ceux retournés.

— **Rappel** : proportion de résultats pertinents extraits par le système.

3. **F-score** : Une mesure combinant précision et rappel.

4.5.2 Méthodologies d'évaluation

1. **Basée sur des ensembles de données standardisés** :
   — Utilisation de jeux de données étiquetés (e.g., ImageNet, COCO).
   — Permet de comparer différents algorithmes sur des bases communes.
2. **Tests utilisateurs** :
   — Observer comment les utilisateurs interagissent avec le système.
   — Mesurer la satisfaction et la convivialité.

4.5.3 Défis d'évaluation

— **Diversité des formats de données** : Les systèmes doivent être évalués sur différents types de multimédias (images, vidéos, sons).
— **Subjectivité de la pertinence** : Les perceptions des utilisateurs varient, ce qui rend difficile une évaluation uniforme.
— **Évolutivité** : Les systèmes doivent être testés à grande échelle pour simuler des conditions réelles.

4.5.4 Approches modernes

— **Mesures de pertinence pondérées** : prendre en compte des scores de pertinence gradués.
— **Apprentissage par renforcement** : optimisation des systèmes en fonction des interactions utilisateur.
— **Benchmarking dynamique** : utiliser des scénarios d'évaluation simulant des cas d'utilisation réels.