perplexity

1 Retrieval Augmented Generation Techniques

D'après l'image fournie, voici les 21 techniques de Retrieval Augmented Generation (RAG) présentées :

1. Simple RAG : Encode le contenu des documents dans un espace vectoriel pour permettre une récupération rapide d'informations pertinentes afin d'améliorer les réponses du modèle.
2. Context Enrichment : Ajoute du contexte environnant à chaque extrait récupéré, améliorant ainsi la cohérence et l'exhaustivité de l'information retournée.
3. Multi-faceted Filtering : Applique diverses techniques de filtrage (métadonnées, seuils de similarité, etc.) pour affiner et améliorer la qualité des résultats récupérés.
4. Fusion Retrieval : Combine la recherche par similarité vectorielle avec la récupération basée sur les mots-clés pour améliorer la récupération de documents.
5. Intelligent Reranking : Réévalue et réorganise les documents initialement récupérés pour s'assurer que l'information la plus pertinente est priorisée pour le traitement ultérieur.
6. Query Transformation : Modifie ou étend la requête originale par réécriture, prompts de retour en arrière et décomposition en sous-requêtes.
7. Hierarchical Indicies : Identifie d'abord les sections pertinentes des documents par des résumés, puis approfondit les détails spécifiques au sein de ces sections.
8. Hypothetical Questions : Transforme les requêtes en documents hypothétiques contenant des réponses, comblant l'écart entre la requête et les distributions de documents dans l'espace vectoriel.
9. Choose Chunk Size : Sélectionne une taille fixe appropriée pour les fragments de texte afin d'équilibrer la préservation du contexte et l'efficacité de la récupération.
10. Semantic Chunking : Contrairement aux méthodes traditionnelles, divise le texte en segments plus significatifs et contextuels.
11. Context Compression : Compresse et extrait les parties les plus pertinentes des documents dans le contexte d'une requête donnée.
12. Explainable Retrieval : Récupère non seulement les documents pertinents, mais fournit également des explications sur la pertinence de chaque document récupéré.
13. Retrieval w/Feedback : Utilise le feedback des utilisateurs sur la pertinence et la qualité des documents récupérés pour affiner les modèles de récupération et de classement.
14. Adaptive Retrieval : Classe les requêtes en différentes catégories et utilise des stratégies de récupération adaptées (factuelles, analytiques, contextuelles, etc.) pour chacune, en tenant compte du contexte et des préférences de l'utilisateur.
15. Iterative Retrieval : Analyse les résultats initiaux et génère des requêtes de suivi pour combler les lacunes ou clarifier l'information.

16. Ensemble Retrieval : Applique différents modèles d'embedding ou algorithmes de récupération et utilise des mécanismes de vote ou de pondération pour déterminer l'ensemble final de documents récupérés.
17. Graph RAG : Récupère des entités et leurs relations à partir d'un graphe de connaissances pertinent pour la requête, combinant avec du texte non structuré pour des réponses plus informatives.
18. Multi-Modal : Intègre des modèles capables de récupérer et de comprendre différentes modalités de données, combinant des informations provenant de textes, d'images et de vidéos.
19. RAPTOR : Utilise la summarisation abstractive pour traiter et résumer récursivement les documents récupérés, organisant l'information dans une structure arborescente pour un contexte hiérarchique.
20. Self RAG : Processus en plusieurs étapes comprenant la décision de récupération, la récupération de documents, l'évaluation de la pertinence, la génération de réponses, et plus encore pour améliorer les réponses du modèle.
21. Corrective RAG : Évalue et corrige dynamiquement le processus de récupération, combinant bases de données vectorielles, recherche web et modèles pour améliorer les réponses.

Ces techniques représentent l'état de l'art dans le domaine du RAG, offrant des moyens sophistiqués d'améliorer la qualité, la pertinence et la précision des informations récupérées pour augmenter les capacités des modèles de génération de langage.

⁂