

Pauta de la presentación

- (5 ptos.) **BONO VIDEO**
- (5 ptos) **PDF enviado antes de presentar**
- (1 pto.) **Contexto:** Contexto del problema que aborda el artículo
- (1 pto.) **Problema de recomendación:** presentar formalmente qué se está recomendando
- (1 pto.) **Contribución:** qué ofrece este artículo comparado con otros trabajos
- (2 pto.) **Estado del arte y marco teórico:** trabajos relacionados y conceptos que se deben conocer para entender el paper
- (4 pto.) **Detalle solución:** detalles, ecuaciones, modelo, etc.
- (3 pto.) **Calidad gráfica y estilo de slides**
- (2 pto.) **Presentación:** modular bien, usar las slides de referencia (no leerlas)
- (3 pto.) **Evaluación:** Presentar los resultados del paper
- (1 pto.) **Referencias:** cita de los paper más importantes
- (1 pto.) **Preguntas:** responder preguntas
- (-x ptos.) **Errores:** descuento por errores
- (1 pto.) **"Puntualidad 15 min":** no pasarse del tiempo para presentar

Evaluación formulario

- **1. Definición y Motivación del Estudio:** La motivación del estudio está claramente definida, incluyendo la importancia del tema y los resultados únicos obtenidos.
- **2. Trabajos Relacionados:** Se presentan estudios previos, destacando sus limitaciones y cómo este paper busca superarlas.
- **3. Métodos Propuestos:** Los métodos del paper están claramente descritos, con una explicación intuitiva de su funcionamiento.
- **4. Resultados:** Se presentan y explican adecuadamente los resultados obtenidos en el estudio.
- **5. Conclusiones y Hallazgos Principales:** Las conclusiones resumen los principales descubrimientos y aportes del estudio.
- **6. Material de Apoyo:** Las diapositivas complementan y facilitan la comprensión del contenido del paper.

Recommender Systems with Generative Retrieval

Eduardo Salinas, Alfonso Badilla y Nicolás Gutiérrez
Grupo 17

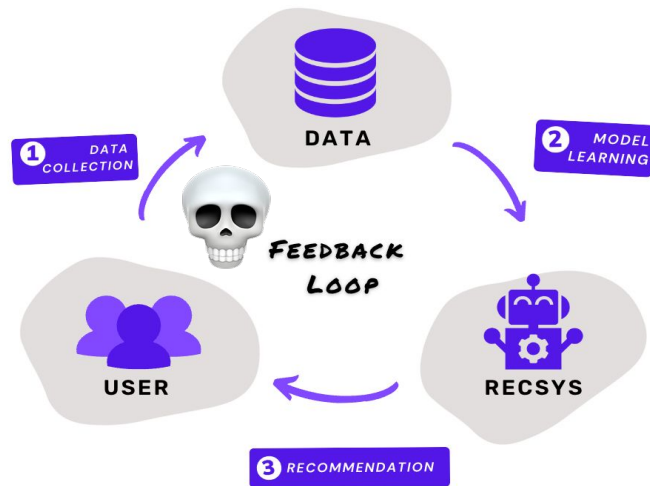
Motivación

- Los sistemas de recomendación modernos depender de **modelos de recuperación y ranking**.
- Desafíos:
 - Generalización limitada con IDs atómicos aleatorios.
 - Alto costo de memoria y falta de adaptabilidad para nuevos ítems.
 - Sesgo de retroalimentación.
- Propuesta:
 - Un nuevo paradigma de recuperación generativa basada en IDs semánticos.
 - Convertir la recuperación en un problema generativo.

ID's Atómicos aleatorios para los ítems:

- Zapatillas Naik:
 - ID = 200192
- Zapatillas Ribuk:
 - ID = 91232
- Zapatos UgoBaus:
 - ID = 312416

Son aleatorios para que se pueda hacer threading para generarlos



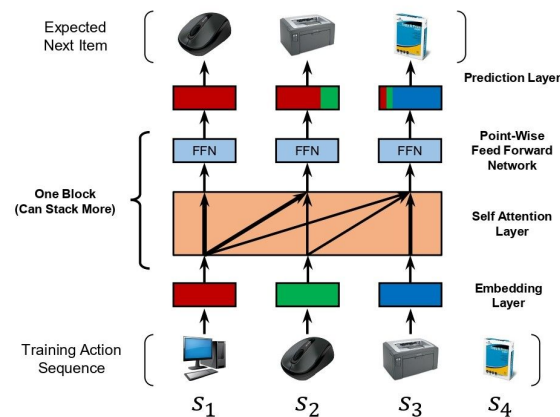
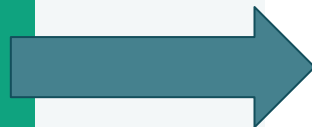
Contexto y problema a resolver

- Sistemas actuales:
 - Métodos como SASRec o BERT4Rec se centran en embeddings densos y autoatención.
 - Persisten problemas de escalabilidad y diversidad en recomendaciones.
- Propuesta **TIGER**:
 - Introducir IDs semánticos tokenizados basados en contenido
 - Utilizar un modelo generativo para predecir secuencias de interacción con mayor precisión

Tokens	Characters
10	27
<div>Tipo:Zapatillas</div> <div>Marca:Ribuk</div>	

Tokens	Characters
10	27
[26620, 25, 57, 11017, 23410, 198, 91809, 163578, 526, 1160]	

Tokenización



Trabajos Relacionados

Modelos secuenciales

- GRU4Rec (Hidasi et al., 2015)
- SASRec (Kang & McAuley, 2018)
- BERT4Rec (Sun et al., 2019)

IDs Semánticos

- VQ-Rec (Hou et al., 2022)
- Hierarchical Semantic IDs (Singh et al., 2018)

Recuperación generativa

- GENRE (Cao et al., 2020)
- DSI (Tay et al., 2022)
- CGR (Lee et al., 2022)

Modelos secuenciales

- Capturan interacciones *user-item* en secuencias
- Usan autoatención y embeddings densos
- Dificultades
 - No manejan bien escenarios de *cold-start*
 - Dependen de índices costosos y estáticos

Trabajos Relacionados

Modelos secuenciales

- GRU4Rec (Hidasi et al., 2015)
- SASRec (Kang & McAuley, 2018)
- BERT4Rec (Sun et al., 2019)

IDs Semánticos

- VQ-Rec (Hou et al., 2022)
- Hierarchical Semantic IDs (Singh et al., 2018)

Recuperación generativa

- GENRE (Cao et al., 2020)
- DSI (Tay et al., 2022)
- CGR (Lee et al., 2022)

IDs Semánticos

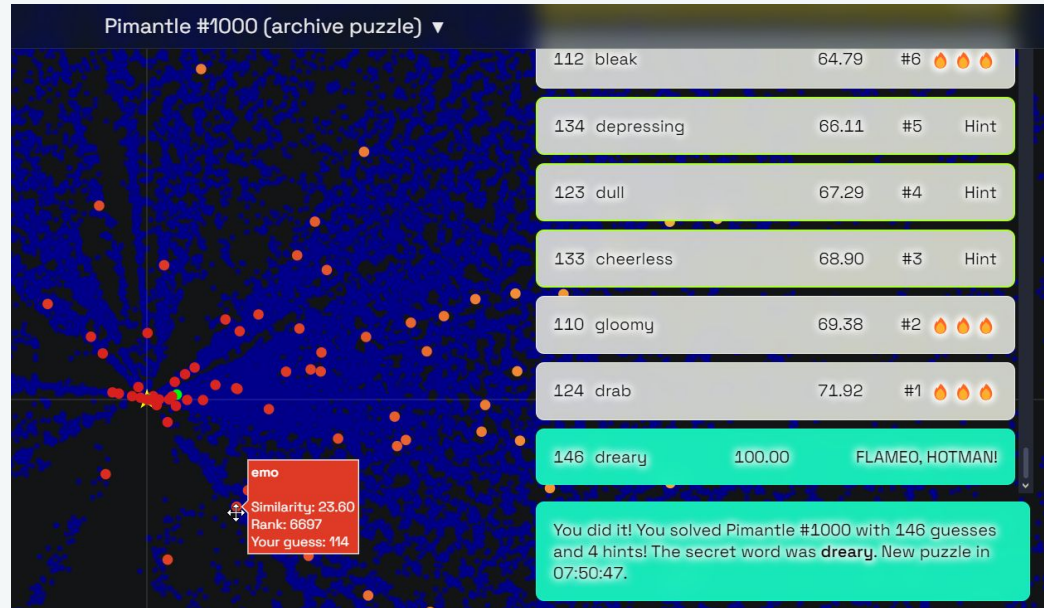
- Generan representaciones significativas basadas en el contenido de los ítems.
- Mejoran la generalización a nuevos ítems
- Dificultades:
 - No son utilizados para predecir ítems de manera generativa
 - Tienen una limitada integración en tareas secuenciales

Convierten palabras exactas a ID's

ID = Zapatillas + Ribuk

ID = Zapatillas + Naik

ID = Zapatos + Ugo Baus



Trabajos Relacionados

Modelos secuenciales

- GRU4Rec (Hidasi et al., 2015)
- SASRec (Kang & McAuley, 2018)
- BERT4Rec (Sun et al., 2019)

IDs Semánticos

- VQ-Rec (Hou et al., 2022)
- Hierarchical Semantic IDs (Singh et al., 2018)

Recuperación generativa

- GENRE (Cao et al., 2020)
- DSI (Tay et al., 2022)
- CGR (Lee et al., 2022)

Recuperación generativa

- Modelos que generan IDs o nombres a documentos directamente
- Eliminan la necesidad de índices tradicionales
- Dificultades:
 - No han sido aplicados a sistemas de recomendación
 - Mayor foco en documentos y ni en interacciones secuenciales

Qué estamos mejorando?

El retrieval generativo es una mejora del retrieval tradicional.

- El retrieval tradicional agrupa los elementos en un espacio vectorial y elige los vecinos más cercanos, como en el juego Pimantle (<https://semantle.pimanrul.es>), para recomendar.
- Requiere coincidencia exacta!!!

baboon

Similarity: 16.80
Rank: 5889
Your guess: 6

45	artery	45.70	#12	Hint
43	aneurysm	46.29	#11	🔥🔥
55	arrhythmia	48.51	#6	🔥🔥🔥
62	arteries	49.00	#5	Hint
59	coronary	49.51	#4	Hint
56	cardiac	51.51	#3	🔥🔥🔥
64	heart	100.00	FLAMEO, HOTMAN!	

You did it! You solved Pimantle #995 with 64 guesses and 6 hints! The secret word was **heart**. New puzzle in 09:45:50.

Share (text)

Share (text+image)

Copy (text)

Copy (image)

Download

Antes convertimos palabras exactas a ID's

De nuevo:

ID = Zapatillas + Ribuk

ID = Zapatillas + Naik

ID = Zapatos + Ugo Baus

Tokens	Characters
10	27
Tipo:Zapatillas	
Marca:Ribuk	

Tokens	Characters
10	27
[26620, 25, 57, 11017, 23410, 198, 91809, 163578, 526, 1160]	

Ahora tokenizamos la información

Tokens

Characters

149

622

El retrieval generativo, qué es lo que mejora?

I met a traveller from an antique land
Who said: Two vast and trunkless legs of stone
Stand in the desert. Near them, on the sand,
Half sunk, a shattered visage lies, whose frown,
And wrinkled lip, and sneer of cold command,
Tell that its sculptor well those passions read
Which yet survive, stamped on these lifeless things,
The hand that mocked them and the heart that fed.
And on the pedestal these words appear:
«My name is Ozymandias, king of kings:
Look on my works, ye Mighty, and despair!»
Nothing beside remains. Round the decay
Of that colossal wreck, boundless and bare
The lone and level sands stretch far away

- El retrieval generativo produce muchos tokens. Luego, pasa todo el universo de tokens por el modelo utilizando ejemplos para aprender a predecir el siguiente token en una secuencia.
- Ya no son palabras, ahora son tokens, así que no necesitamos coincidencia exacta.
- Sobre conjuntos de tokens podemos usar transformers igual que antes.
- Gracias tokenizador de ChatGPT.

Text

Token IDs

Entonces nos quedan los siguientes ítems:

Tipo:Zapatillas [26620, 25, 57, 11017, 23410, 198, 91809, 163578, 526, 1160]

Marca:Ribuk

Tipo:Zapatillas [26620, 25, 57, 11017, 23410, 198, 91809, 25, 11398, 507]

Marca:Naik

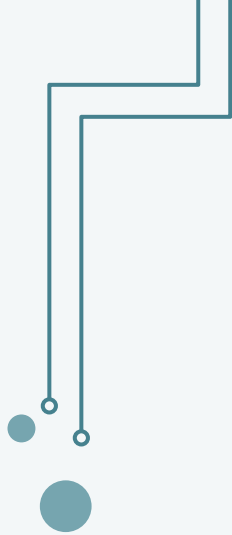
Tipo:Zapatos [26620, 25, 139991, 10322, 198, 91809, 25, 52, 2319, 33, 8674]

Marca:UgoBaus

Luego, esos tokens se agrupan en tokens más grandes

- ejemplo: un token para “Tipo:Zapat...”)
- ejemplo2: si hubiera un ítem de “Marca:PipeBaus” se podría hacer un token de “Baus”)

La idea es que identifiquen ítems diferentes. Son a la vez identificador y localizador en el espacio vectorial de ítems.



Ahora que tengo tokens excelentes,
puedo aplicar un equivalente a SVD

**(Singular Value Decomposition, el de
separar en vectores latentes)**



Y qué pasa si hay colisión?

Si te sucede que dos ítems tienen los mismos tokens (por un problema de parámetros, o simplemente que son ítems muy parecidos), simplemente agregas ID's

Handling Collisions. Depending on the distribution of semantic embeddings, the choice of codebook size, and the length of codewords, semantic collisions can occur (*i.e.*, multiple items can map to the same Semantic ID). To remove the collisions, we append an extra token at the end of the ordered semantic codes to make them unique. For example, if two items share the Semantic ID (12, 24, 52), we append additional tokens to differentiate them, representing the two items as (12, 24, 52, 0) and (12, 24, 52, 1). To detect collisions, we maintain a lookup table that maps Semantic IDs to corresponding items. Note that collision detection and fixing is done only once after the RQ-VAE model is trained. Furthermore, since Semantic IDs are integer tuples, the lookup table is efficient in terms of storage in comparison to high dimensional embeddings.

(literalmente eso dice el artículo)

Finalmente lo que estoy haciendo es:

Ítem: → Token: → ID:

Tipo: Zapatillas
Marca: Ribuk

Tipo:Zapatillas
Marca:Ribuk

Menos ids si usáramos buenos parámetros

[26620, 25, 57, 11017, 23410, 198, 91809, 163578, 526, 1

Experimentos

Dataset utilizado: Amazon Product Reviews

Experimentos

Dataset utilizado: Amazon Product Reviews

- Reviews de usuarios
- Metadata de ítems
- Mayo 1996 - Julio 2014

Experimentos

Dataset utilizado: Amazon Product Reviews

- Reviews de usuario
- Metadata de ítems
- Mayo 1996 - Julio 2014

Categorías usadas

- Beauty
- Sports and Outdoors
- Toys and Games

Experimentos

Dataset utilizado: Amazon Product Reviews

- Reviews de usuario
- Metadata de ítems
- Mayo 1996 - Julio 2014

Categorías usadas

- Beauty
- Sports and Outdoors
- Toys and Games

Table 6: Dataset statistics for the three real-world benchmarks.

Dataset	# Users	# Items	Sequence Length	
			Mean	Median
Beauty	22,363	12,101	8.87	6
Sports and Outdoors	35,598	18,357	8.32	6
Toys and Games	19,412	11,924	8.63	6

Evaluación

Métricas usadas:

- Recall@K
- NDCG@K

Evaluación

Métricas usadas:

- Recall@K

$$recall@k = \frac{\text{number of recommended relevant items among top } k}{\text{number of all relevant items in the system}}$$

- NDCG@K

Evaluación

Métricas usadas:

- Recall@K

$$recall@k = \frac{\text{number of recommended relevant items among top } k}{\text{number of all relevant items in the system}}$$

- NDCG@K

$$NDCG@K = \frac{DCG@K}{IDCG@K} = \frac{\sum_{i=1}^{k \text{ (actual order)}} \frac{Gains}{\log_2(i+1)}}{\sum_{i=1}^{k \text{ (ideal order)}} \frac{Gains}{\log_2(i+1)}}$$

Evaluación

Table 1: Performance comparison on sequential recommendation.

Methods	Beauty			
	Recall @5	NDCG @5	Recall @10	NDCG @10
P5 [8]	0.0163	0.0107	0.0254	0.0136
Caser [33]	0.0205	0.0131	0.0347	0.0176
HGN [25]	0.0325	0.0206	0.0512	0.0266
GRU4Rec [11]	0.0164	0.0099	0.0283	0.0137
BERT4Rec [32]	0.0203	0.0124	0.0347	0.0170
FDSA [42]	0.0267	0.0163	0.0407	0.0208
SASRec [17]	0.0387	0.0249	0.0605	0.0318
S ³ -Rec [44]	0.0387	0.0244	0.0647	0.0327
TIGER [Ours]	0.0454	0.0321	0.0648	0.0384
	+17.31%	+29.04%	+0.15%	+17.43%

Evaluación

Table 1: Performance comparison on sequential recommendation.

Methods	Toys and Games			
	Recall @5	NDCG @5	Recall @10	NDCG @10
P5 [8]	0.0070	0.0050	0.0121	0.0066
Caser [33]	0.0166	0.0107	0.0270	0.0141
HGN [25]	0.0321	0.0221	0.0497	0.0277
GRU4Rec [11]	0.0097	0.0059	0.0176	0.0084
BERT4Rec [32]	0.0116	0.0071	0.0203	0.0099
FDSA [42]	0.0228	0.0140	0.0381	0.0189
SASRec [17]	0.0463	0.0306	0.0675	0.0374
S ³ -Rec [44]	0.0443	0.0294	0.0700	0.0376
TIGER [Ours]	0.0521	0.0371	0.0712	0.0432
	+12.53%	+21.24%	+1.71%	+14.97%

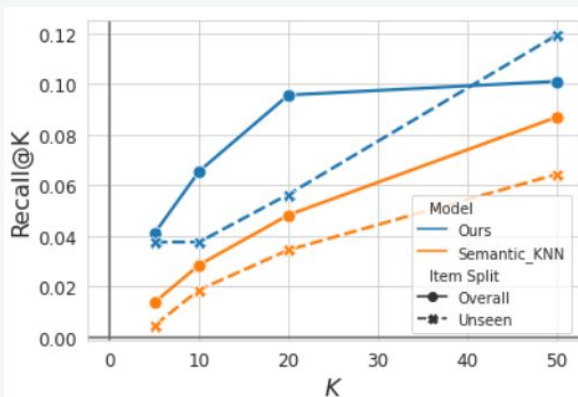
Evaluación

Table 1: Performance comparison on sequential recommendation.

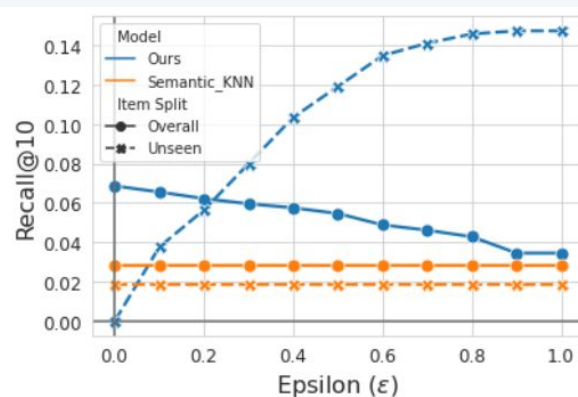
Methods	Sports and Outdoors			
	Recall @5	NDCG @5	Recall @10	NDCG @10
P5 [8]	0.0061	0.0041	0.0095	0.0052
Caser [33]	0.0116	0.0072	0.0194	0.0097
HGN [25]	0.0189	0.0120	0.0313	0.0159
GRU4Rec [11]	0.0129	0.0086	0.0204	0.0110
BERT4Rec [32]	0.0115	0.0075	0.0191	0.0099
FDSA [42]	0.0182	0.0122	0.0288	0.0156
SASRec [17]	0.0233	0.0154	0.0350	0.0192
S ³ -Rec [44]	<u>0.0251</u>	<u>0.0161</u>	<u>0.0385</u>	<u>0.0204</u>
TIGER [Ours]	0.0264	0.0181	0.0400	0.0225
	+5.22%	+12.55%	+3.90%	+10.29%

Métricas Relevantes

Problema de Cold-Start Recommendation



(a) Recall@K vs. K, ($\epsilon = 0.1$).



(b) Recall@10 vs. ϵ .

Figure 5: Performance in the cold-start retrieval setting.

Métricas Relevantes

Diversidad en la recomendación

Temperature	Entropy@10	Entropy@20	Entropy@50
T = 1.0	0.76	1.14	1.70
T = 1.5	1.14	1.52	2.06
T = 2.0	1.38	1.76	2.28

Table 3: The entropy of the category distribution predicted by the model for the Beauty dataset.

Métricas Relevantes

Capas en sequence-to-sequence model

Number of Layers	Recall@5	NDCG@5	Recall@10	NDCG@10
3	0.04499	0.03062	0.06699	0.03768
4	0.0454	0.0321	0.0648	0.0384
5	0.04633	0.03206	0.06596	0.03834

Table 5: Recall and NDCG metrics for different number layers.

Métricas Relevantes

Efecto de agregar información de los usuarios

	Recall@5	NDCG@5	Recall@10	NDCG@10
No user information	0.04458	0.0302	0.06479	0.0367
With user id (reported in the paper)	0.0454	0.0321	0.0648	0.0384

Table 8: The effect of providing user information to the recommender system

Posible Problema

Predicción de IDs inválidos

Posible Problema

Predicción de IDs inválidos

- IDs semánticos de largo 4

Posible Problema

Predicción de IDs inválidos

- IDs semánticos de largo 4
- ID no lleva a ningún item en el dataset

Posible Problema

Predicción de IDs inválidos

- IDs semánticos de largo 4
- ID no lleva a ningún item en el dataset

Problema en la práctica

- ~ 0.1% - 1.6% IDs inválidos recomendados
- A pesar del problema, TIGER muestra un desempeño mejorado

Conclusión

El framework TIGER presenta resultados interesantes:

Conclusión

El framework TIGER presenta resultados interesantes:

- Diversidad

Conclusión

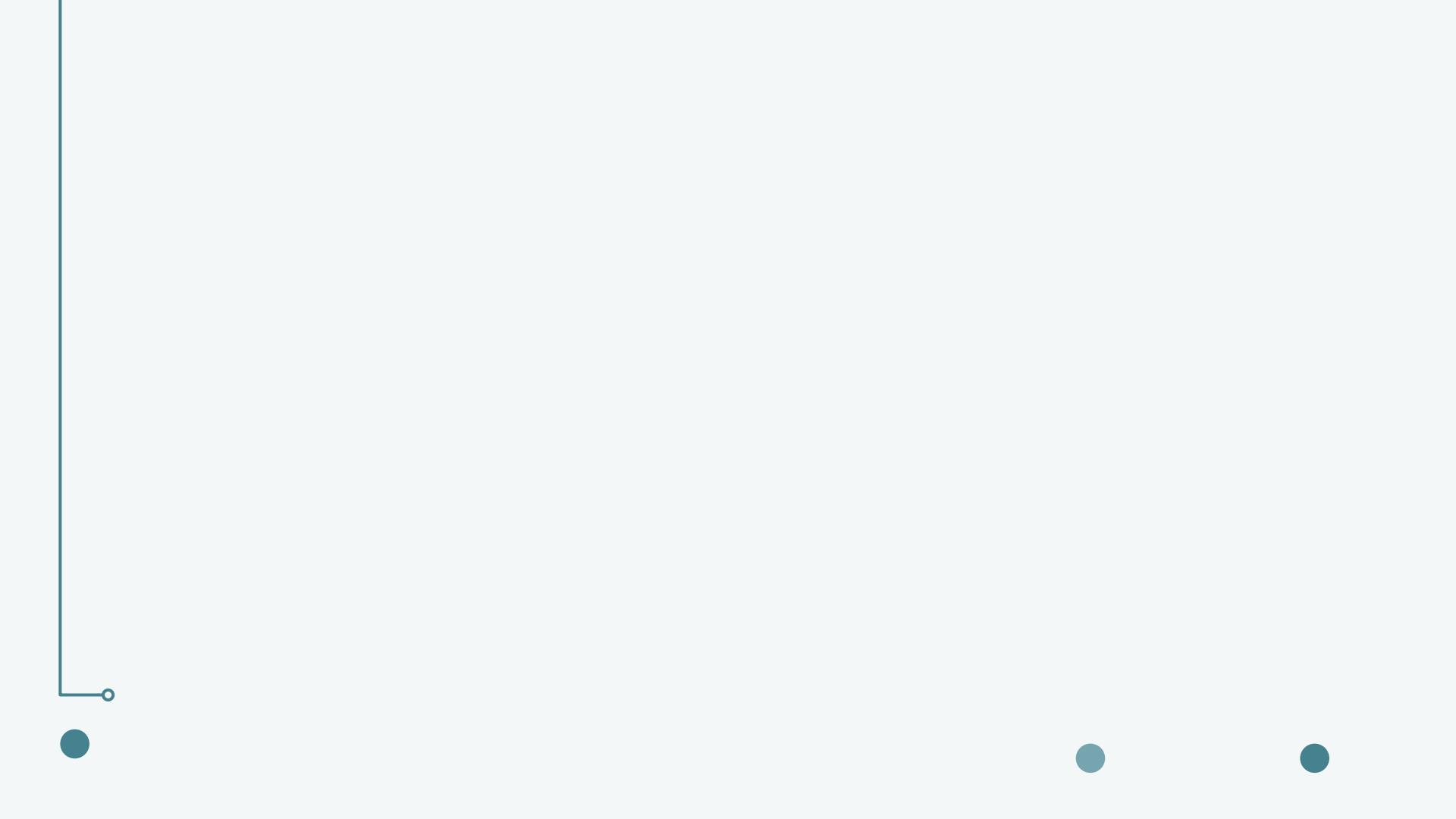
El framework TIGER presenta resultados interesantes:

- Diversidad
- Cold-Start

Conclusión

El framework TIGER presenta resultados interesantes:

- Diversidad
- Cold-Start
- Performance



Referencias

Acá no sé bien qué hacer, pero creo que hay que poner los papers citados por el artículo.

Dice la pauta que citar los papers más importantes, entonces esos que explican por ejemplo la creación de los semantic ID deberíamos agregar acá.

