Universidade de São Paulo - USP
Instituto de Ciências Matemáticas e de Computação - ICMC
Programa de Pós-Graduação em Ciências da Computação e Matemática Computacional

Ambiente de Data Warehouse Para Imagens Médicas Baseado Em Similaridade

Luis Marcelo Bortolotti

O imageDWE

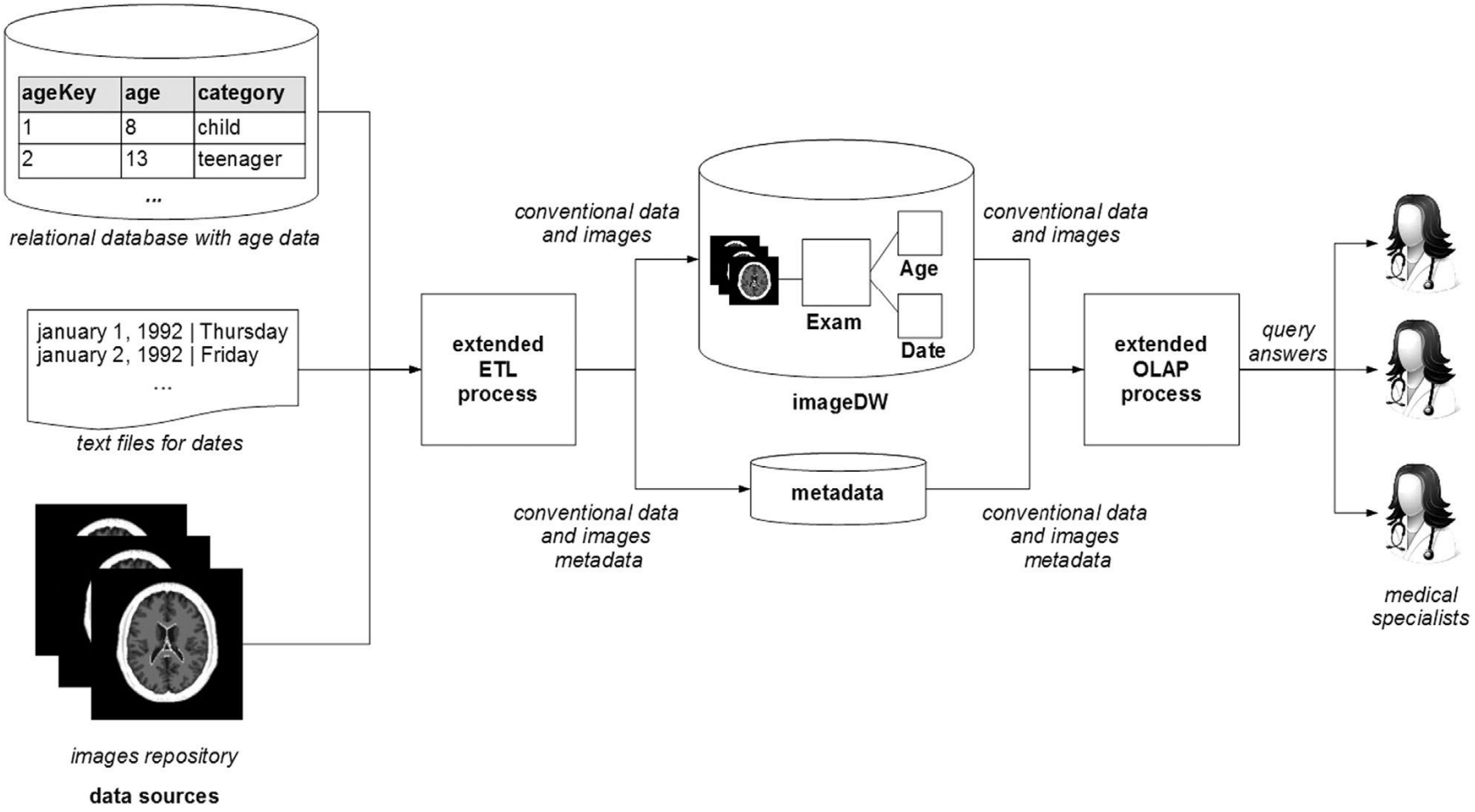
imageDWE

- **imageDWE**, proposto por Teixeira et al. (2015)
  - Armazenamento em conjunto de dados convencionais e de dados imagens
    - Extensão do esquema estrela para englobar tabelas dimensão projetadas para armazenar dados de imagens
  - Diminuição do gap semântico com o uso do conceito de camadas perceptuais (*perceptual layers*)

imageDWE

- **imageDWE**, proposto por Teixeira et al. (2015)
  - Extensão do processo de ETL, para suportar imagens. São gerados dados para as camadas perceptuais armazenadas no imageDW
  - Extensão do processamento de consultas OLAP, integrando a execução de consultas por similaridade às consultas OLAP convencionais
  - Introdução de uma técnica de indexação, que contempla a especificação de um índice e a definição de diferentes estratégias de processamento de consultas

O *imageDWE*



Fonte: Teixeira et al. (2015)

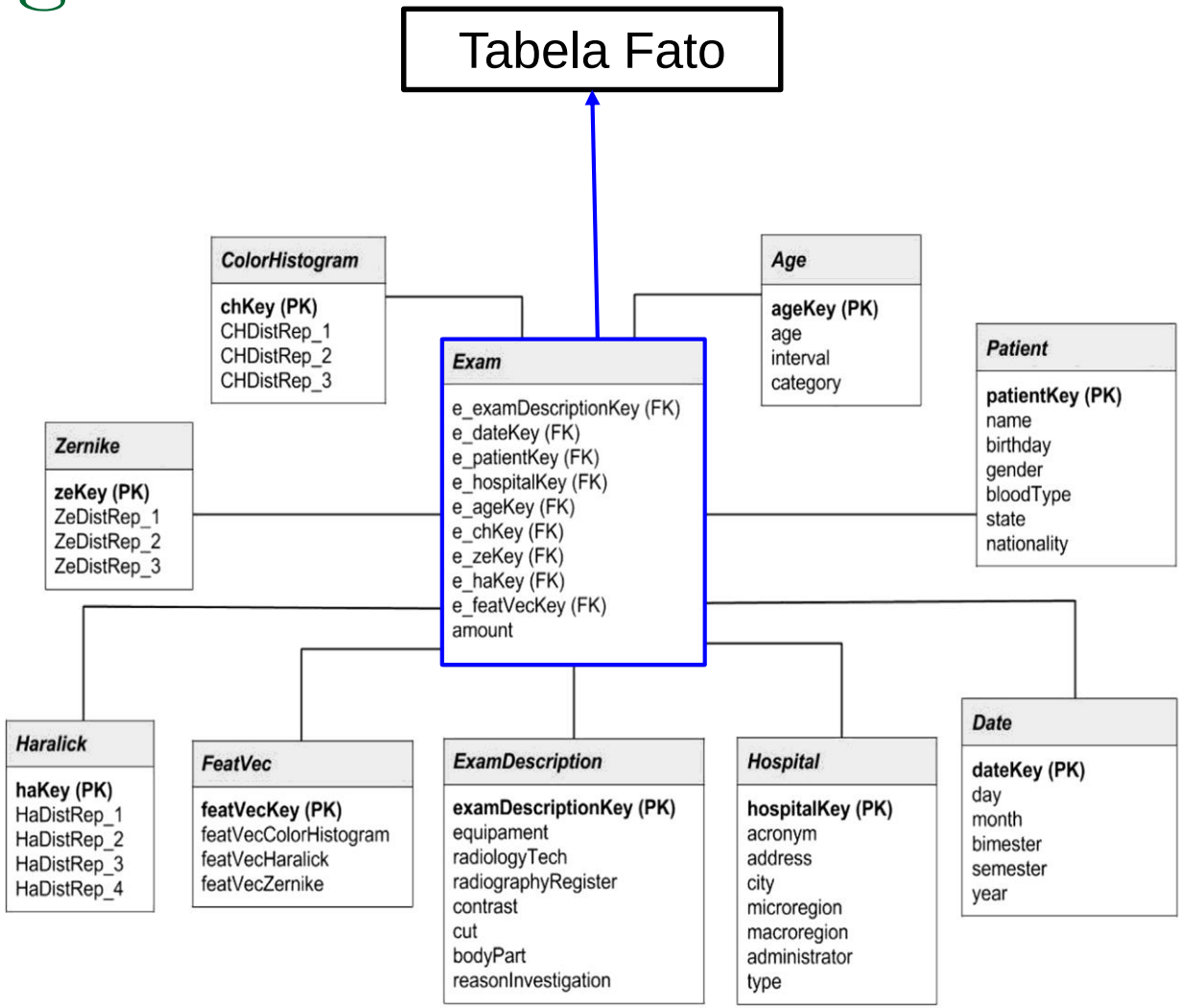
O *imageDW*

- O *imageDW* é modelado segundo o esquema estrela
  - **Tabela fato –** contém medidas numéricas e chaves estrangeiras para as tabelas dimensão; mantém a relação entre cada imagem e seus dados convencionais
  - **Tabelas dimensão convencionais –** contém a chave primária e vários atributos descritivos
  - **Tabelas dimensão de imagens –** contém os dados intrínsecos das imagens e suportam a definição de várias camadas perceptuais

Camada Perceptual e Tabelas Dimensão de Imagens

- Uma **Camada Perceptual** é a representação de conjunto de imagens de acordo com um descritor de características de forma a permitir a busca por similaridade
- As imagens são representadas por:
  - Seus vetores de características, gerados de acordo com um descritor de características específico e;
  - Seus dados de similaridade, geradas de acordo com um espaço métrico específico

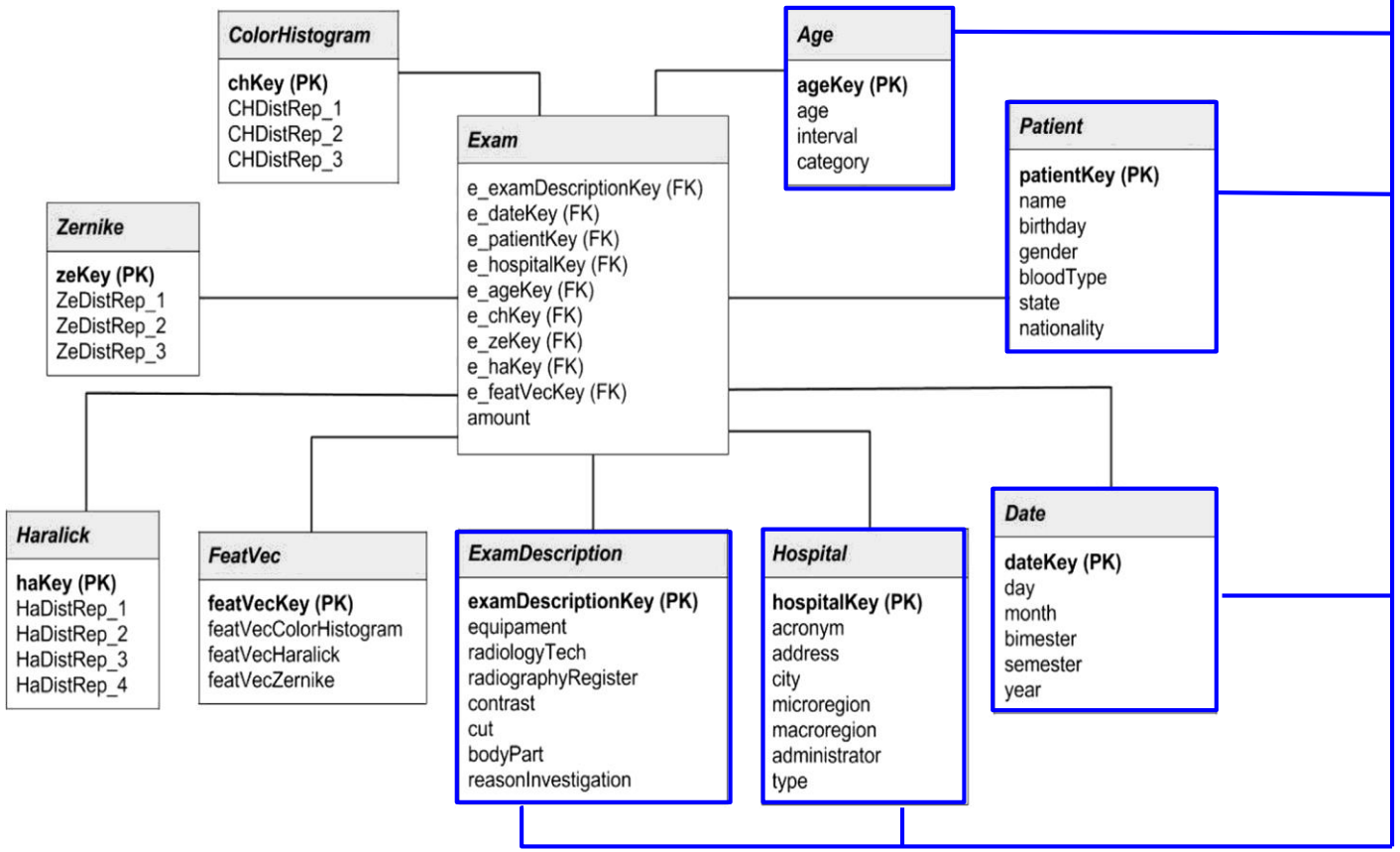
O *imageDW*



Fonte: Teixeira et al. (2015)

O *imageDW*

Tabelas Dimensão Convencionais

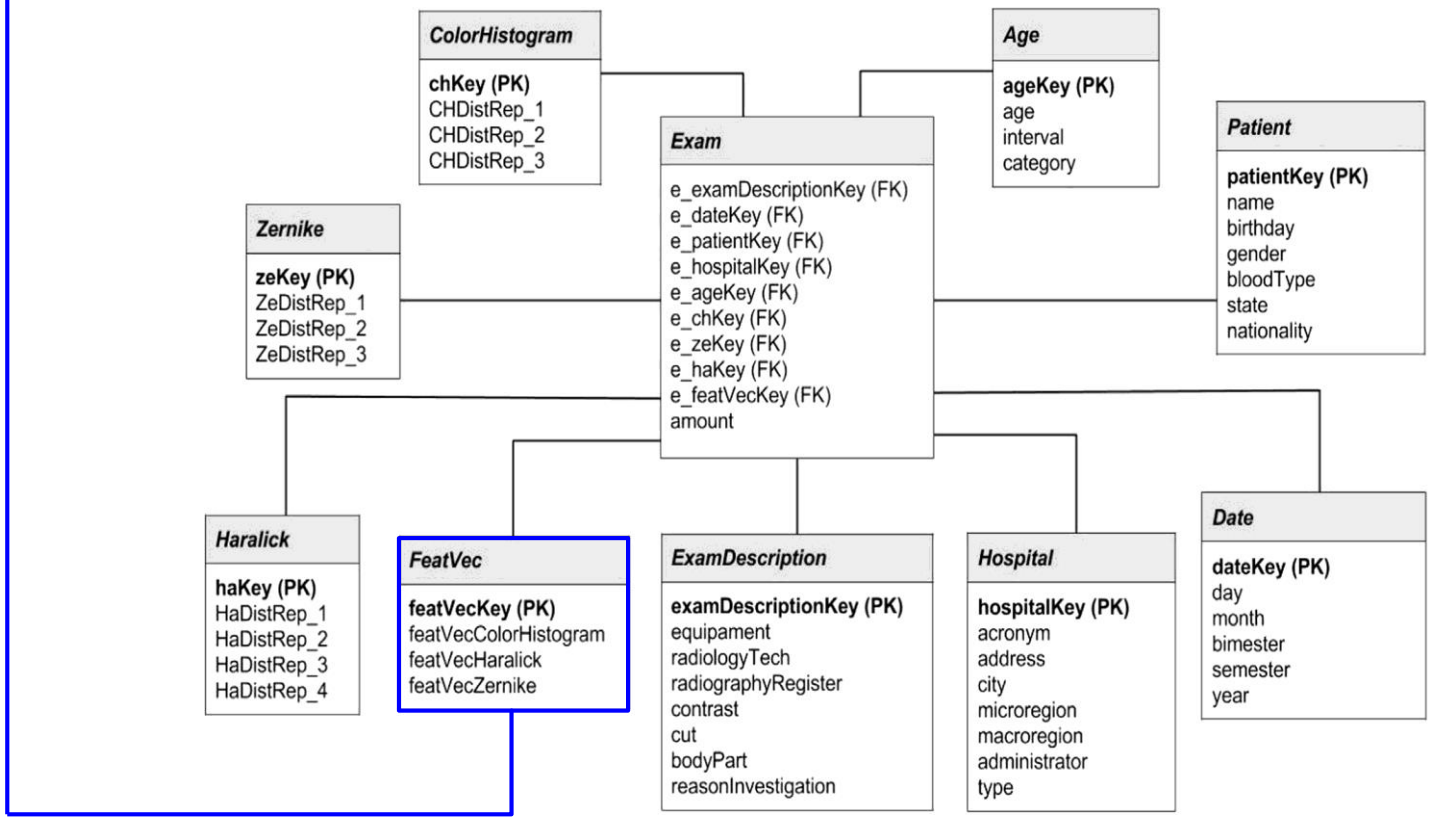


Fonte: Teixeira et al. (2015)

O *imageDW*

Tabela Dimensão de Imagem Primária

vetores de características

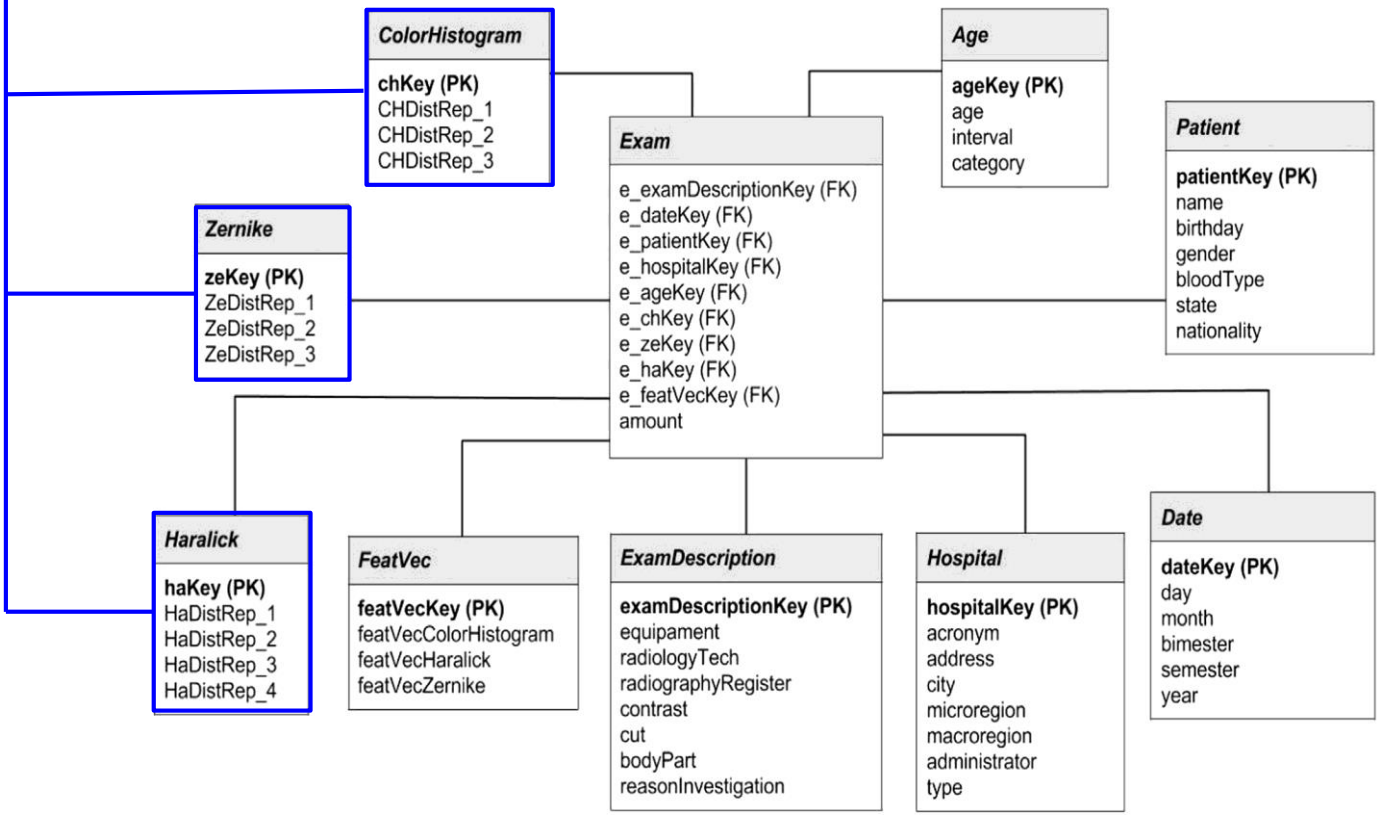


Fonte: Teixeira et al. (2015)

O *imageDW*

Tabelas Dimensão de Imagem Secundárias

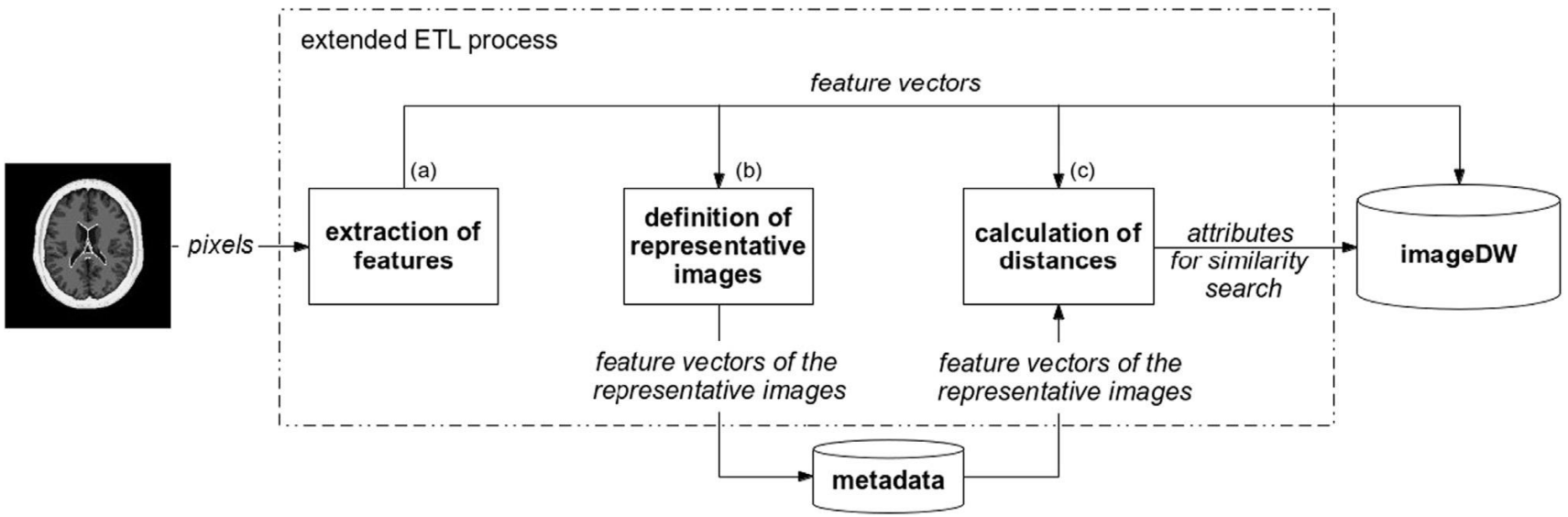
atributos para busca por similaridade (distâncias de uma imagem até as representativas, de acordo com a técnica Omni)



Fonte: Teixeira et al. (2015)

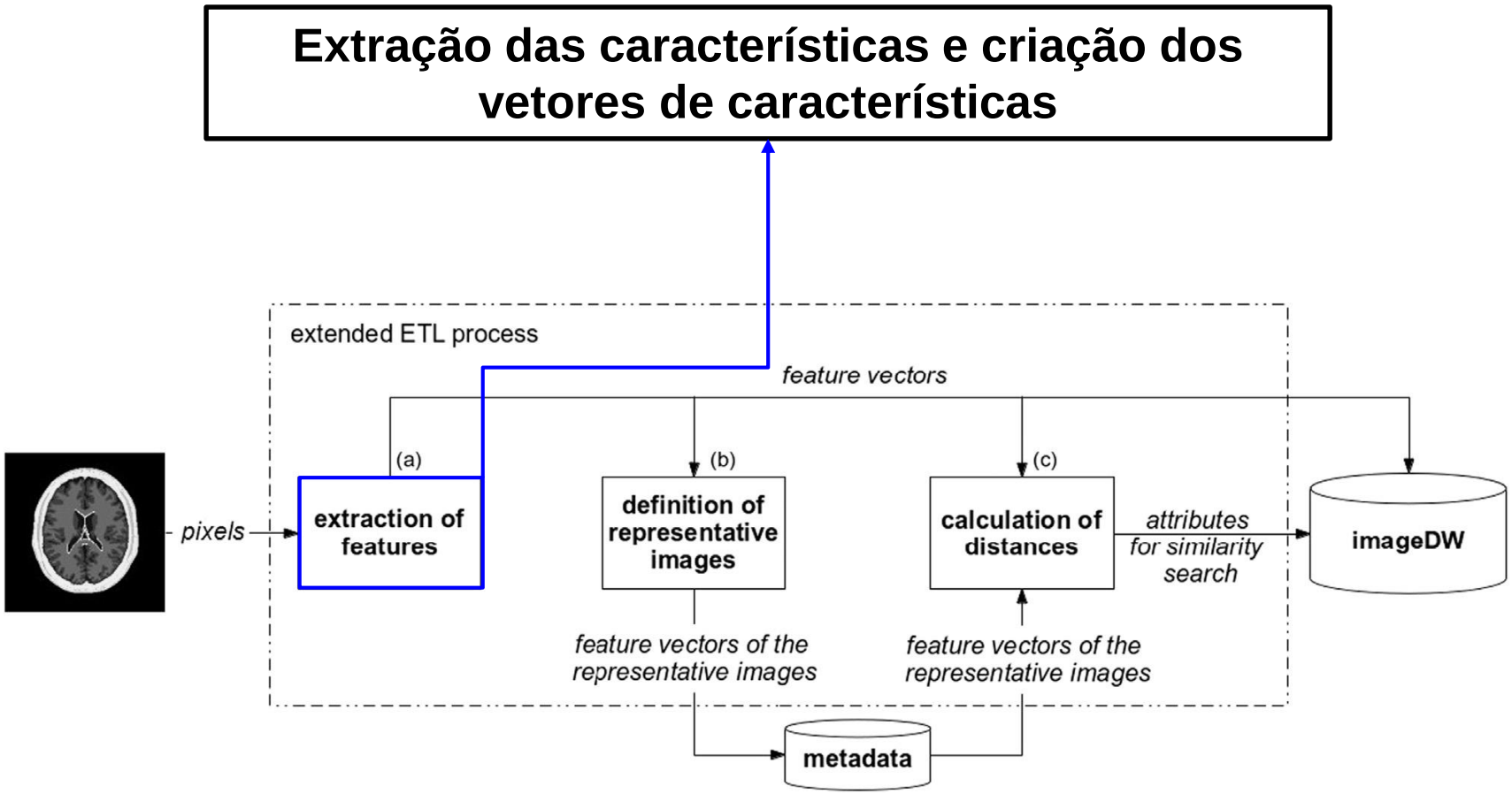
O Processo ETL Estendido

- Extração e armazenamento de características de imagens
- As etapas se repetem para cada camada perceptual



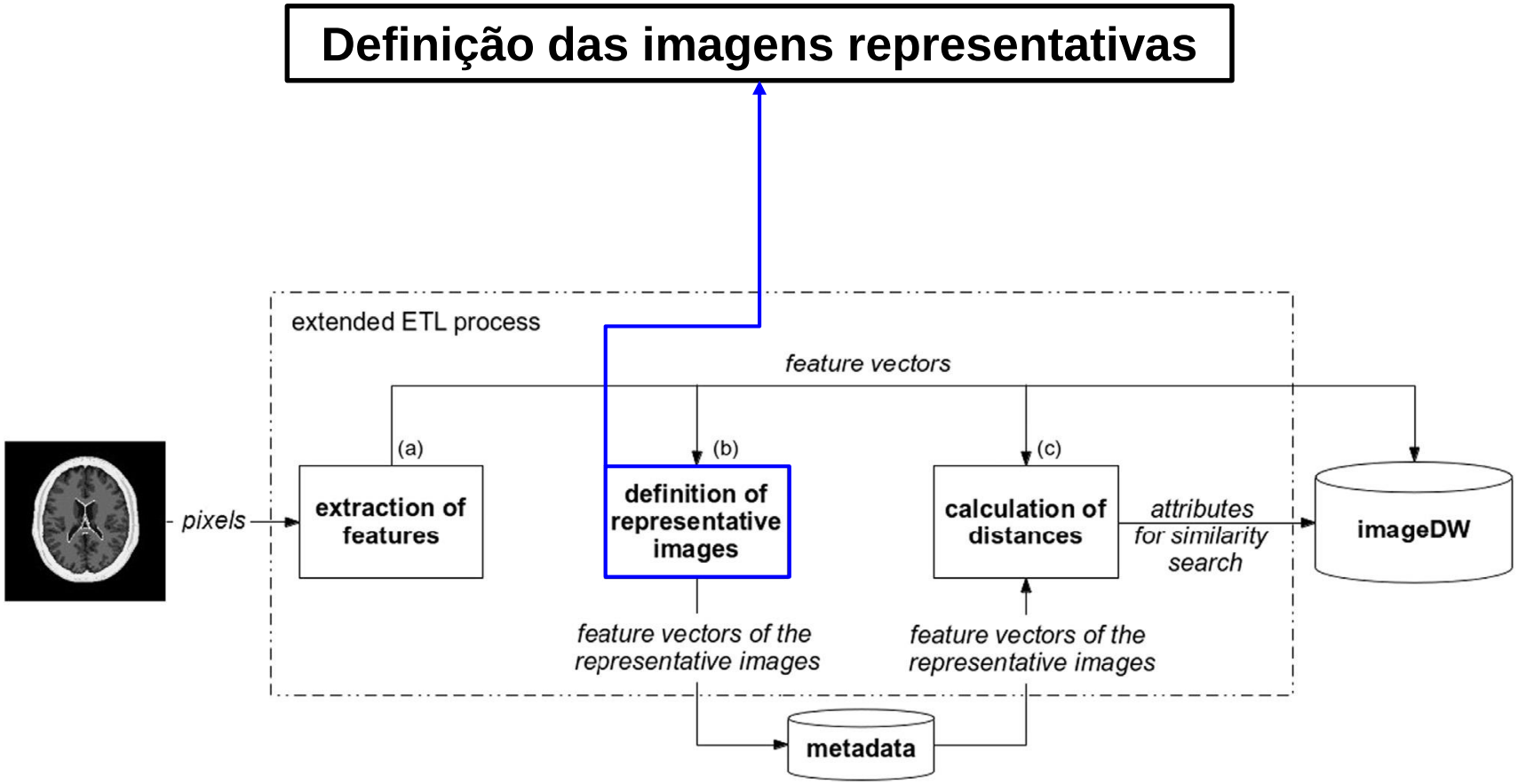
Fonte: Teixeira et al. (2015)

O Processo ETL Estendido



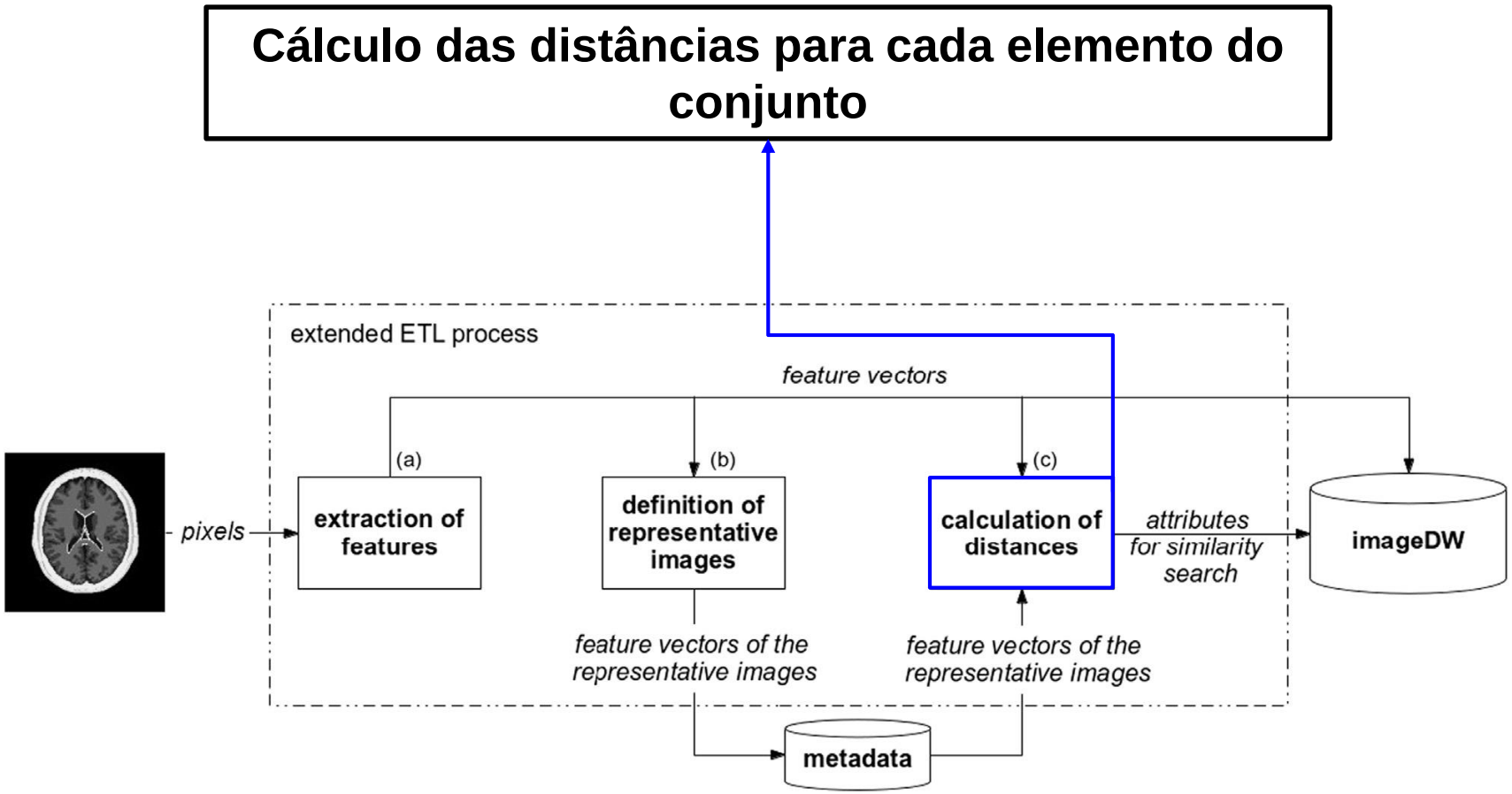
Fonte: Teixeira et al. (2015)

O Processo ETL Estendido



Fonte: Teixeira et al. (2015)

O Processo ETL Estendido



Fonte: Teixeira et al. (2015)

O Processamento de Consultas OLAP Estendido

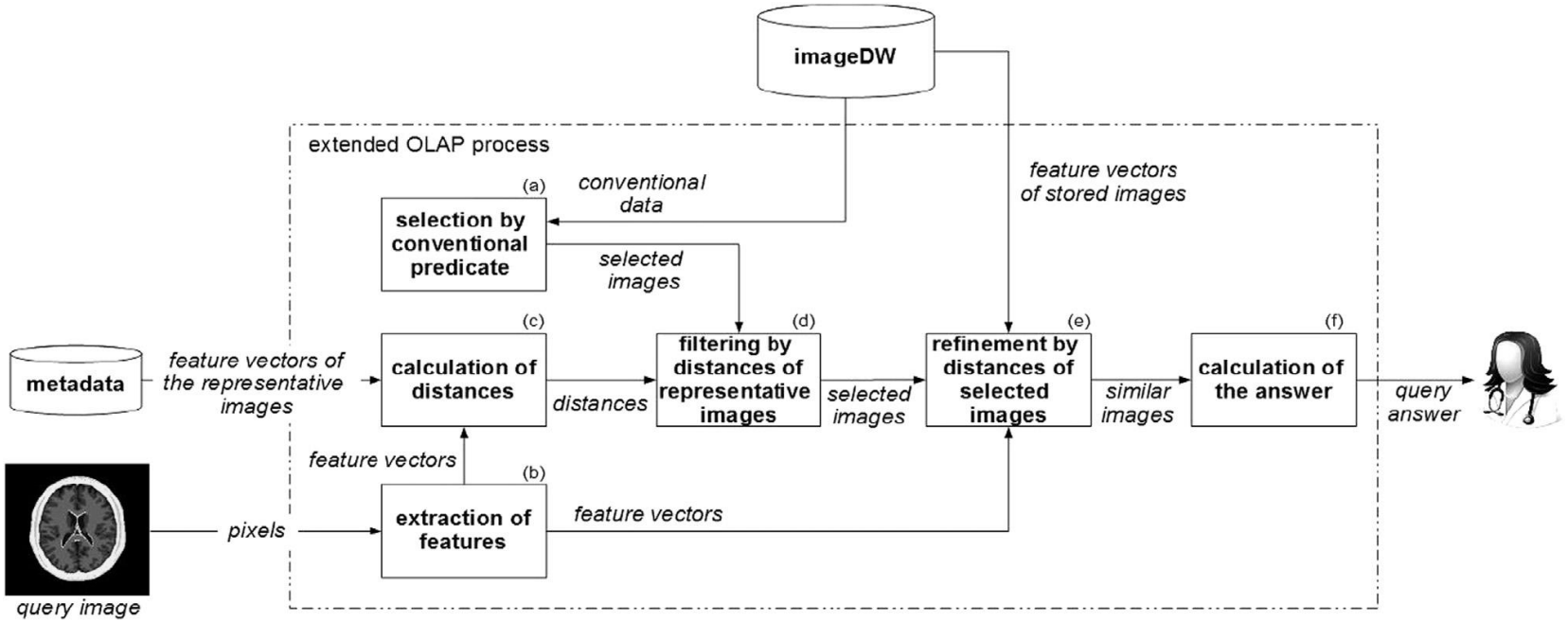
- Extensão das consultas OLAP, incorporando a execução de consultas por similaridade
  - Atributos convencionais são comparados pelos operadores relacionais tradicionais
  - Os predicados de busca por similaridade estão associados às características intrínsecas das imagens, dando suporte a uma ou mais camadas perceptuais

O Processamento de Consultas OLAP Estendido

- Divididas em seis etapas:
  a. Seleção via predicados convencionais
  b. Extração de características
  c. Cálculo das distâncias
  d. Filtro pelas distâncias até as imagens representativas
  e. Refinamento pelas distâncias até imagens selecionadas
  f. Cálculo da resposta

O Processamento de Consultas OLAP Estendido

- Acontece em 6 etapas

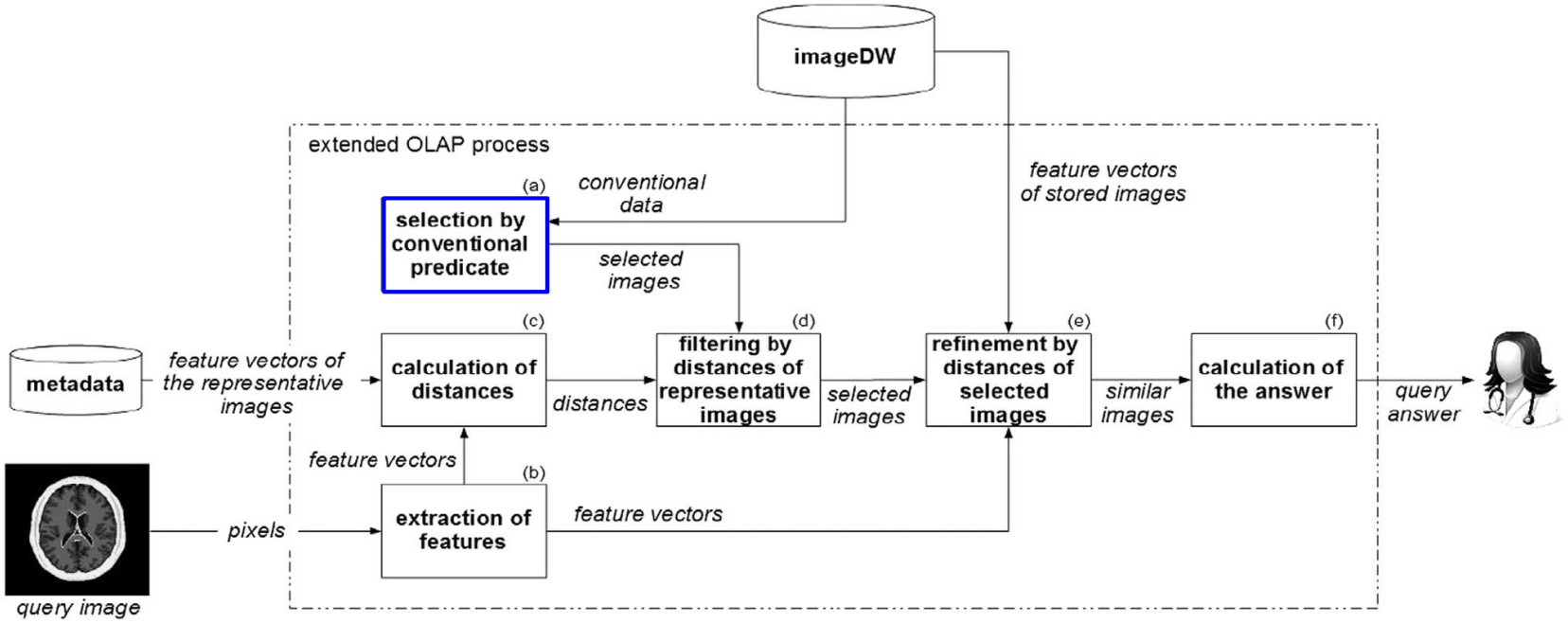


Fonte: Teixeira et al. (2015)

O Processamento de Consultas OLAP Estendido

(a) Seleção via predicados convencionais

Filtra o conjunto de resultados possíveis com base em dados alfanuméricos tradicionais

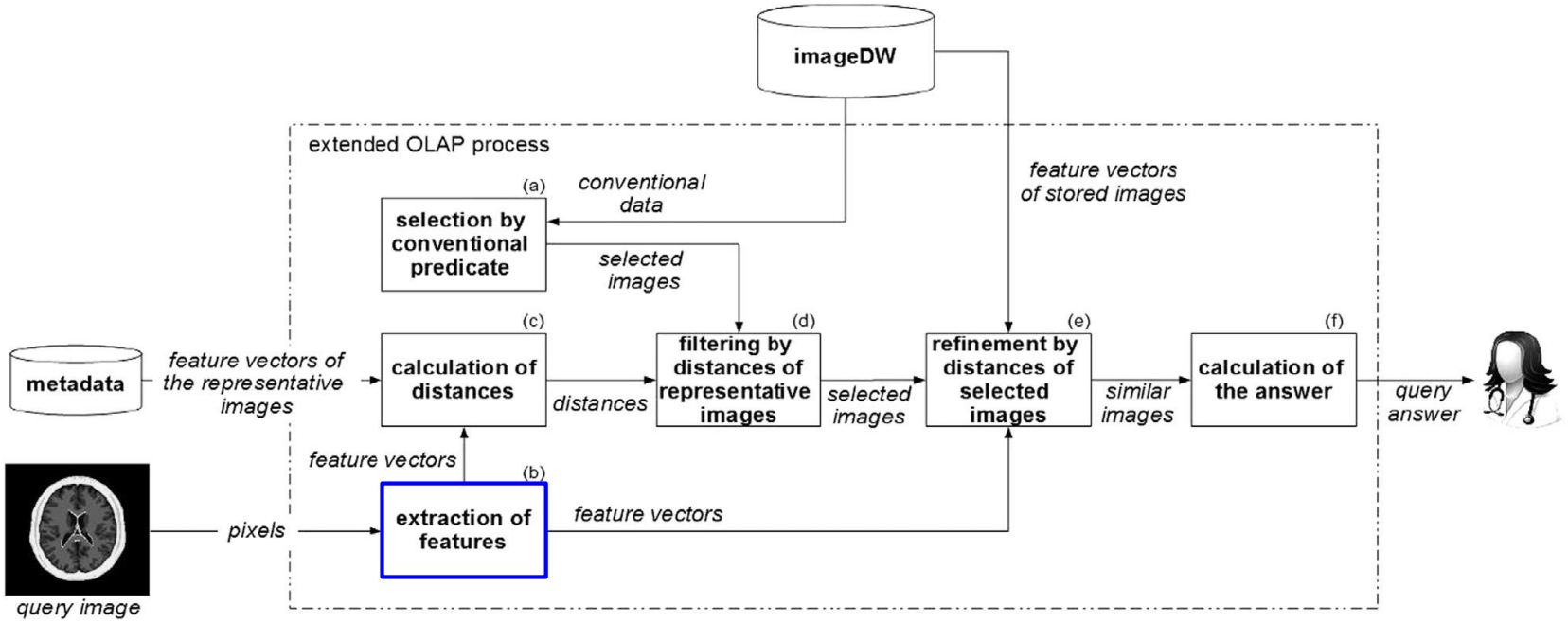


Fonte: Teixeira et al. (2015)

O Processamento de Consultas OLAP Estendido

(b) Extração de características

Extrai as características da imagem de busca de acordo com os descritores utilizados nas camadas perceptuais

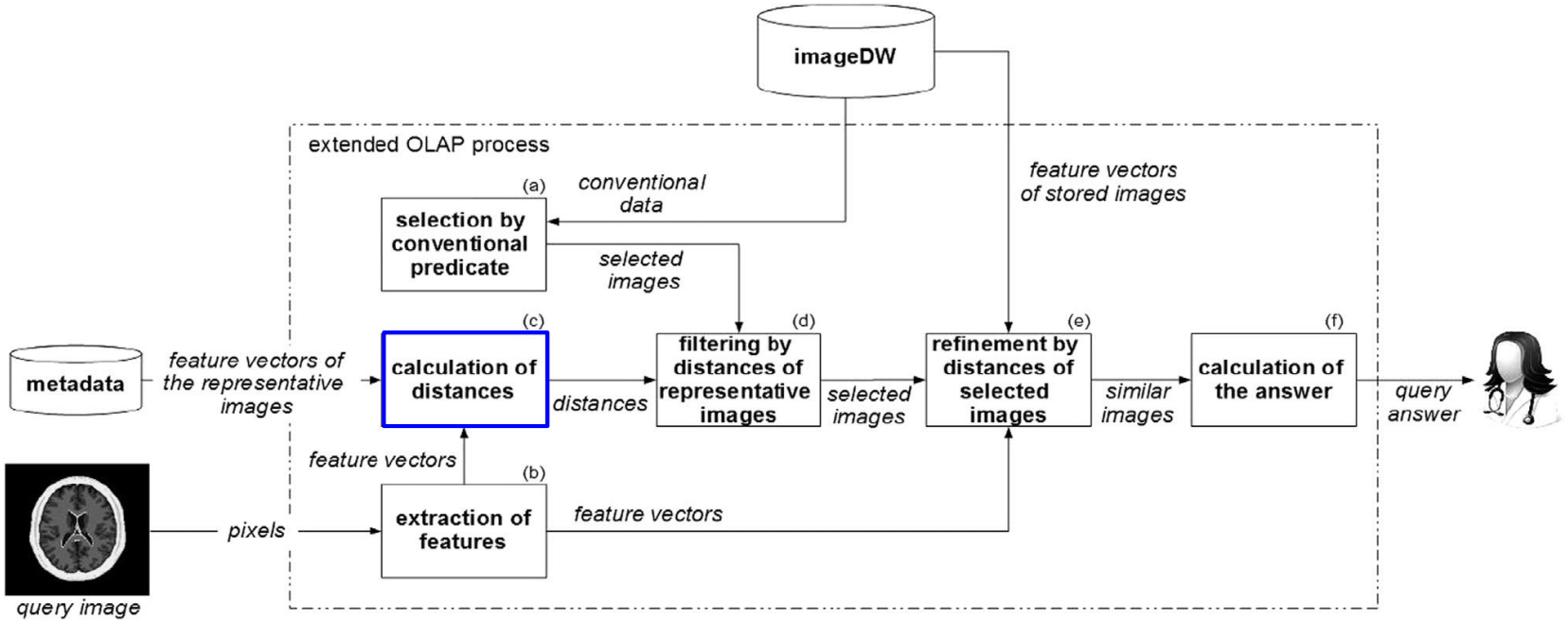


Fonte: Teixeira et al. (2015)

O Processamento de Consultas OLAP Estendido

(c) Cálculo das distâncias

Calcula a distância da imagem de busca até as imagens representativas de cada camada perceptual, utilizando a mesma função distância do processo ETL estendido

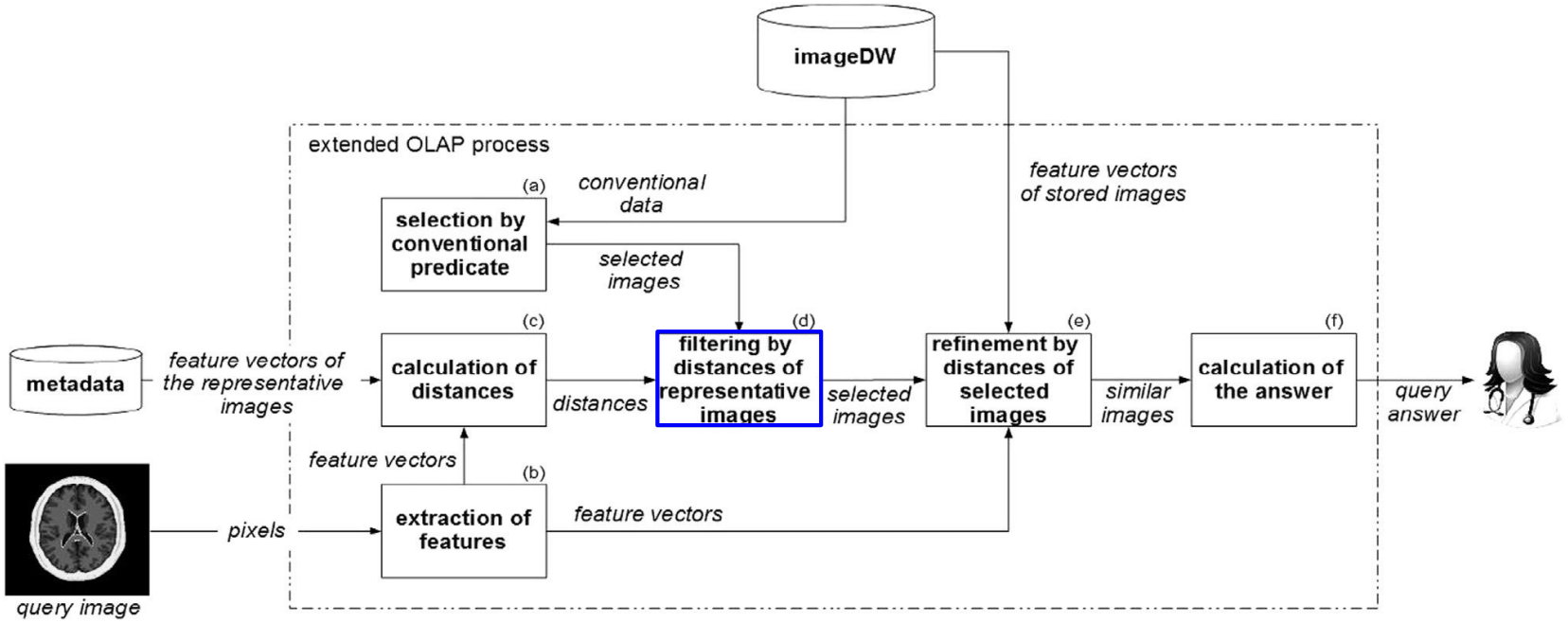


Fonte: Teixeira et al. (2015)

O Processamento de Consultas OLAP Estendido

(d) Filtro pelas distâncias até as imagens representativas

*As imagens selecionadas na etapa **(a)** são filtradas de acordo com os predicados de busca por similaridade*

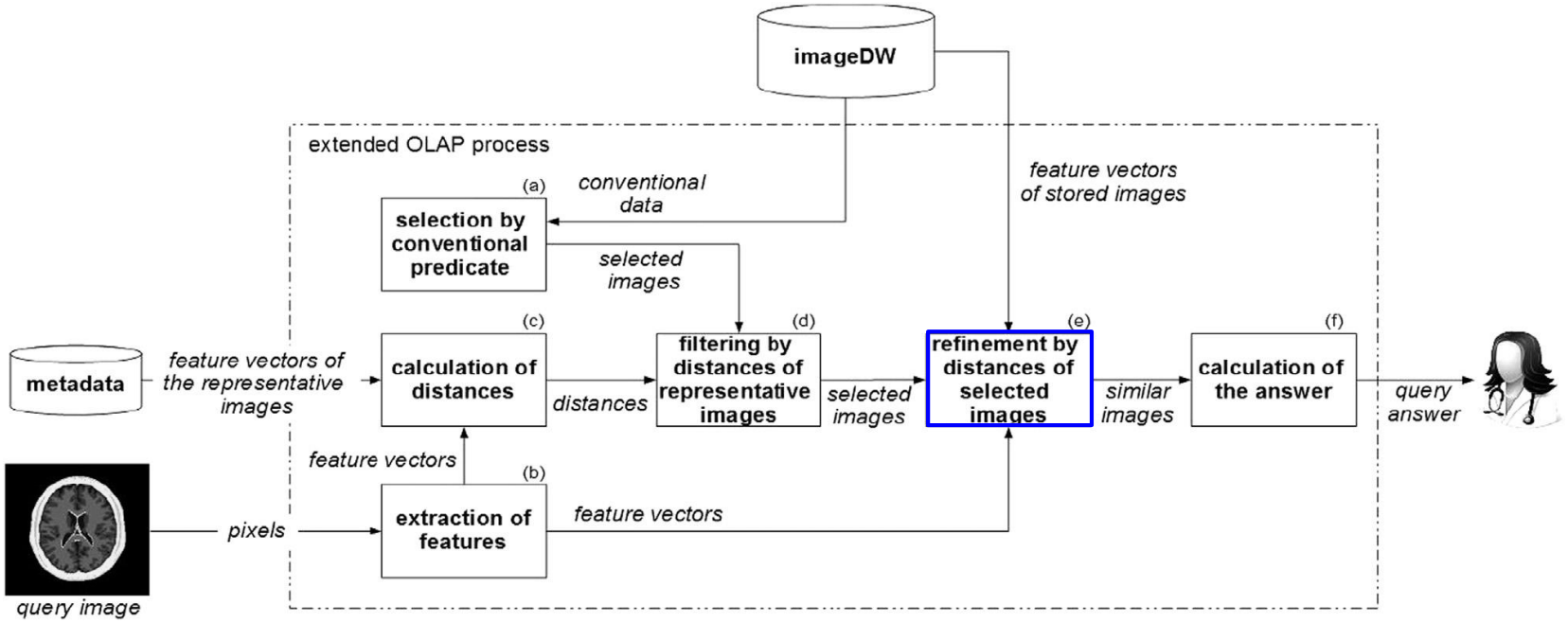


Fonte: Teixeira et al. (2015)

O Processamento de Consultas OLAP Estendido

(e) Refinamento pelas distâncias até imagens selecionadas

*Com base nas distâncias entre a imagem de busca e as imagens resultantes da etapa **(d)**; são eliminados os falsos positivos*

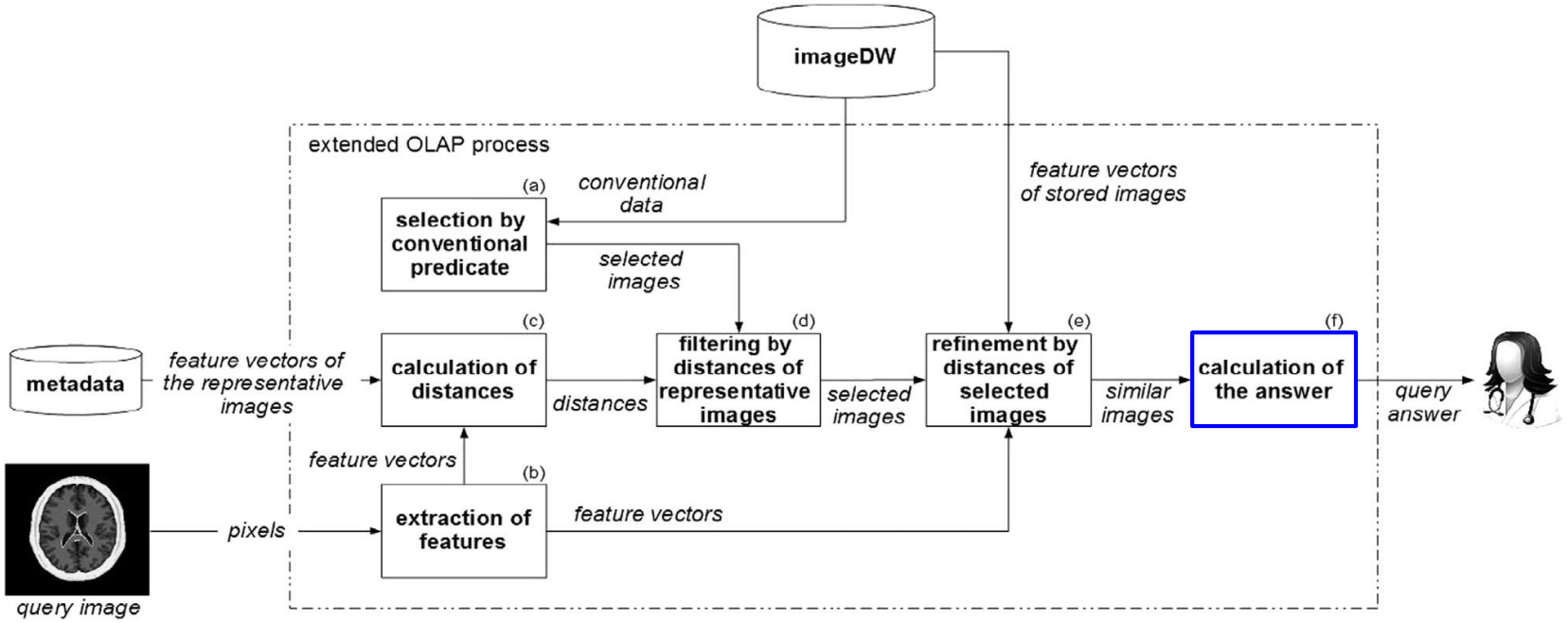


Fonte: Teixeira et al. (2015)

O Processamento de Consultas OLAP Estendido

(f) Cálculo da resposta

*Usa as imagens similares encontradas na etapa **(e)**, para calcular a resposta*



Fonte: Teixeira et al. (2015)

O Processamento de Consultas OLAP Estendido

- **Exemplo:** consultas OLAP estendidas
  - Considere a consulta definida por um especialista

”Quantas imagens são similares a uma dada imagem de mamografia (*imagem de consulta*), de acordo com as camadas perceptuais de *Histograma de Cores* (*ColorHistogram*) e *Zernike* e dentro do raio de busca r_q, que ocorreram no estado de Nova York, na faixa etária de 30 a 40 anos, nos anos de 1993 a 1994, nos hospitais da macrorregião do Estado de Nova York”

O Processamento de Consultas OLAP Estendido

a. Seleção via predicados convencionais
   - As imagens são filtradas de acordo com o ano em que foram geradas, com a macrorregião onde estão localizados os hospitais, com o estado onde os pacientes estão e, sua idade na datas em que foram realizados os exames

b. Extração de características
   - São gerados dois vetores de características, um para o Histograma de Cores e outro para Zernike

c. Cálculo das distâncias
   - São calculadas três distâncias para as imagens representativas da camada perceptual de Histograma de Cores, e outras três para a camada perceptual de Zernike

O Processamento de Consultas OLAP Estendido

d. Filtro pelas distâncias até as imagens representativas
   - Com as distâncias calculadas na etapa **(c)** e um raio de busca r_q, calcular a *mbOr* da consulta para cada camada perceptual; filtrar apenas as imagens presentes em todas as mbOr

e. Refinamento pelas distâncias até imagens selecionadas
   - São eliminados os falsos positivos

f. Cálculo da resposta
   - Usa funções de agregação para determinar a quantidade de imagens retornadas

O *imageSJBindex*

- É um índice *star-join bitmap* que indexa a distância entre as imagens do conjunto de dados e as imagens representativas, de acordo com cada camada perceptual
- É uma matriz *m x n* onde *m* é o número de vetores, de forma que exista pelo menos um vetor para cada imagem representativa de cada camada perceptual, e *n* é a quantidade de tuplas indexadas
- Utiliza ***binning***
  - Cada vetor binário representa um intervalo de valores, ao invés de valores discretos

O *imageSJBindex*

- **Exemplo:** *imageSJBindex* para o esquema estrela do imageDW

# tuple	Color Histogram									Zernike									Haralick
	CHDistRep_1			CHDistRep_2			CHDistRep_3			ZeDistRep_1			ZeDistRep_2			ZeDistRep_3			…
	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	…
0	0	1	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	…
1	1	0	0	0	1	0	0	0	1	1	0	0	0	1	0	0	0	1	…
2	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	…
…	…			…			…			…			…			…			…

Fonte: Teixeira et al. (2015)

O *imageSJBindex*

- O *imageSJBindex* modifica a etapa **(d)** do Processamento de Consultas OLAP Estendido: *Filtro pelas distâncias até as imagens representativas*
  - São indexadas as distâncias entre as imagens armazenadas e as imagens representativas
  - Não há necessidade de acessar o imageDW
  - As imagens pertencentes à mbOr através de operações bit a bit

O *imageSJBindex*

- Uso do *imageSJBindex*
  - Considere o predicado de busca por similaridade do exemplo 3: ”Quantas imagens são similares a uma dada imagem de mamografia (*imagem de consulta*), de acordo com as camadas perceptuais de *Histograma de Cores* (*ColorHistogram*) e *Zernike* e dentro do raio de busca r_q=1”
    - As distâncias calculadas para os representantes de cada camada perceptual (o intervalo é dado por r_q):

Representative image	Distance	Interval
CHDistRep_1	2	[1,3]
CHDistRep_2	5	[4,6]
CHDistRep_3	8	[7,9]

Distância entre a imagem de busca e as imagens representativas da camada perceptual do Histograma de Cores

Representative image	Distance	Interval
ZeDistRep_1	1	[0,2]
ZeDistRep_2	4	[3,5]
ZeDistRep_3	7	[6,8]

Distância entre a imagem de busca e as imagens representativas da camada perceptual do Zernike

Fonte: Teixeira et al. (2015)

O *imageSJBindex*

- Uso do *imageSJBindex*
  - Primeiro, na etapa **(d)** melhorada é feita a intersecção do *imageSJBindex* com os intervalos calculados da imagem de busca

# tuple	Color Histogram									Zernike									Haralick
	CHDistRep_1			CHDistRep_2			CHDistRep_3			ZeDistRep_1			ZeDistRep_2			ZeDistRep_3			...
	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	...
0	0	1	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	...
1	1	0	0	0	1	0	0	0	1	1	0	0	0	1	0	0	0	1	...
2	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	...
...	...			...			...			...			...			...			...

Representative image	Distance	Interval
CHDistRep_1	2	[1,3]

CHDistRep_1		
[0,2]	[3,5]	[6,8]
0	1	0
1	0	0
1	0	0

Fonte: Teixeira et al. (2015)

O *imageSJBindex*

- Uso do *imageSJBindex*
  - Primeiro, na etapa **(d)** melhorada é feita a intersecção do *imageSJBindex* com os intervalos calculados da imagem de busca

Imagem de busca

Representative image	Distance	Interval
CHDistRep_1	2	[1,3]
CHDistRep_2	5	[4,6]
CHDistRep_3	8	[7,9]

Representative image	Distance	Interval
ZeDistRep_1	1	[0,2]
ZeDistRep_2	4	[3,5]
ZeDistRep_3	7	[6,8]

# tuple	Color Histogram									Zernike									Haralick
	CHDistRep_1			CHDistRep_2			CHDistRep_3			ZeDistRep_1			ZeDistRep_2			ZeDistRep_3			…
	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	[0,2]	[3,5]	[6,8]	…
0	0	1	0	0	1	0	0	0	1	1	0	0	0	0	1	1	0	0	…
1	1	0	0	0	1	0	0	0	1	1	0	0	0	1	0	0	0	1	…
2	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	…
…	…			…			…			…			…			…			…

Fonte: Teixeira et al. (2015)

O *imageSJBindex*

- Uso do *imageSJBindex*
  - Primeiro, na etapa **(d)** melhorada é feita a intersecção do *imageSJBindex* com os intervalos calculados da imagem de busca

Imagem de busca

Representative image	Distance	Interval
CHDistRep_1	2	[1,3]
CHDistRep_2	5	[4,6]
CHDistRep_3	8	[7,9]

Representative image	Distance	Interval
ZeDistRep_1	1	[0,2]
ZeDistRep_2	4	[3,5]
ZeDistRep_3	7	[6,8]

	Color Histogram					Zernike		
	CHDistRep_1		CHDistRep_2		CHDistRep_3	ZeDistRep_1	ZeDistRep_2	ZeDistRep_3
# tuples	[0,2]	[3,5]	[3,5]	[6,8]	[6,8]	[0,2]	[3,5]	[6,8]
0	0	1	1	0	1	1	0	0
1	1	0	1	0	1	1	1	1
2	1	0	0	1	0	0	0	1

Fonte: Teixeira et al. (2015)

O *imageSJBindex*

- Uso do *imageSJBindex*
  - Como são selecionados dois *bins* para *CHDISTRep_1* e dois para *CHDISTRep_2*, é realizada uma operação OR entre os *bins*

CHDistRep_1				
[0,2]		[3,5]		OR result
0		1	=	1
1	OR	0	=	1
1		0	=	1

CHDistRep_2				
[3,5]		[6,8]		OR result
1		0	=	1
1	OR	0	=	1
0		1	=	1

  - Por fim, é aplicado uma operação AND entre os bins e o resultado da etapa **(d)** melhorada é obtido

	Color Histogram			Zernike			
# tuples	CHDistRep_1	CHDistRep_2	CHDistRep_3	ZeDistRep_1	ZeDistRep_2	ZeDistRep_3	AND result
0	1	1	1	1	0	0	= 0
1	1	1	1	1	1	1	= 1
2	1	1	0	0	0	1	= 0

Fonte: Teixeira et al. (2015)

Estratégias de Processamento de Consultas

Estratégias de Processamento de Consultas

- A estratégias de processamento de consultas visam usar o *imageSJBindex* no processamento de consultas OLAP por similaridade

- As estratégias são baseadas em dois aspectos:
  - I - indexação opcional utilizando o *imageSJBindex* dos dados convencionais e vetores de características
  - II - Análise em diferentes ordens dos predicados de busca convencionais e de similaridade usando o *imageSJBindex*

I - indexação opcional utilizando o *imageSJBindex*

- Determina se os vetores de características serão indexados ou não utilizando o *imageSJBindex*
- Quando os vetores de características são indexados não existe a necessidade de acessar o *imageDW* para o processamento de consultas
  - Utilizadas apenas em vetores de características de baixa dimensionalidade, devido aos altos custos envolvidos no processamento de consultas com dados de alta dimensionalidade

II - Análise em diferentes ordens dos predicados de busca

- Utilizando o *imageSJBindex*, são definidas duas abordagens para determinar qual a ordem de análise dos predicados de busca convencionais e de similaridade: **dividida (*split*)** e **combinada (*combined*)**

II - Análise em diferentes ordens dos predicados de busca

- Na abordagem **dividida**, primeiro são filtradas as imagens utilizando os predicados de similaridade usando o *imageSJBindex* e, depois filtra as imagens utilizando os predicados convencionais
  - Apresenta melhor suporte para consultas utilizando apenas os predicados de similaridade

II - Análise em diferentes ordens dos predicados de busca

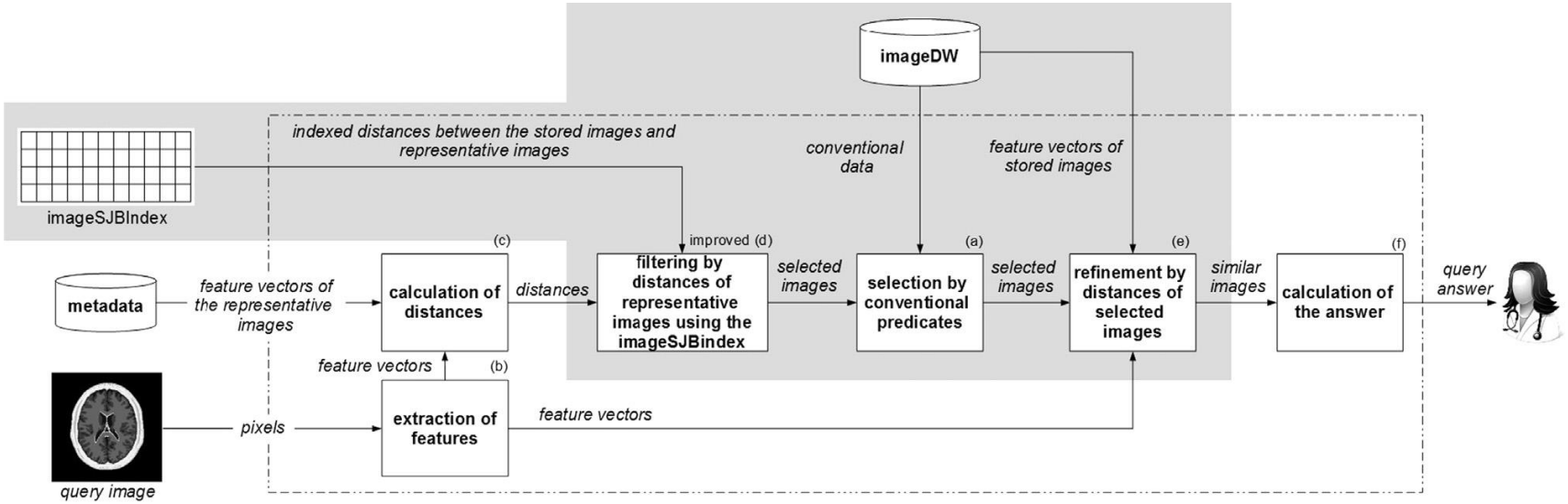
- Na abordagem **combinada**, os predicados convencionais e de similaridades são filtrados juntos, utilizando o *imageSJBindex*
  - Ambos predicados devem estar indexados pelo *imageSJBindex*
  - Maior eficiência em consultas que utilizam ambos os predicados convencionais e de similaridade

Estratégias de Processamento de Consultas

- A combinação das duas abordagens de processamento de predicados, é possível determinar 4 estratégias diferente de processamento de consultas
  - *splitNotIndexFV*
  - *splitIndexFV*
  - *combinedNotIndexFV*
  - *combinedIndexFV*

splitNotIndexFV

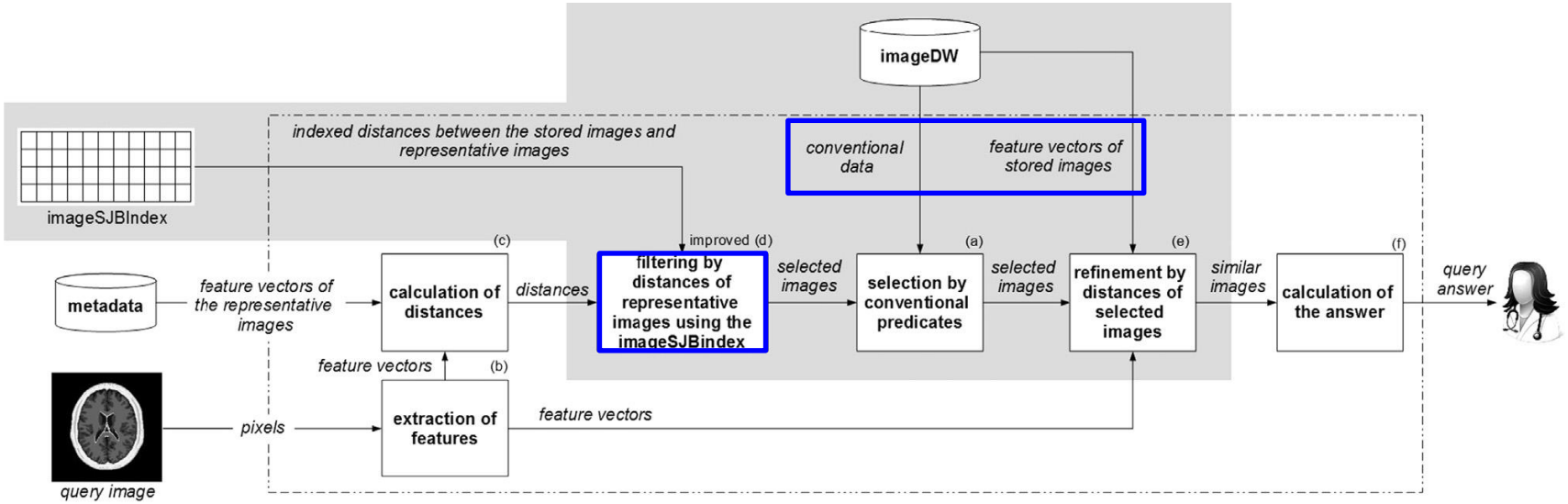
- É a junção das abordagens **dividida (*split*)** com a de **não indexar os vetores de características**



Fonte: Teixeira et al. (2015)

splitNotIndexFV

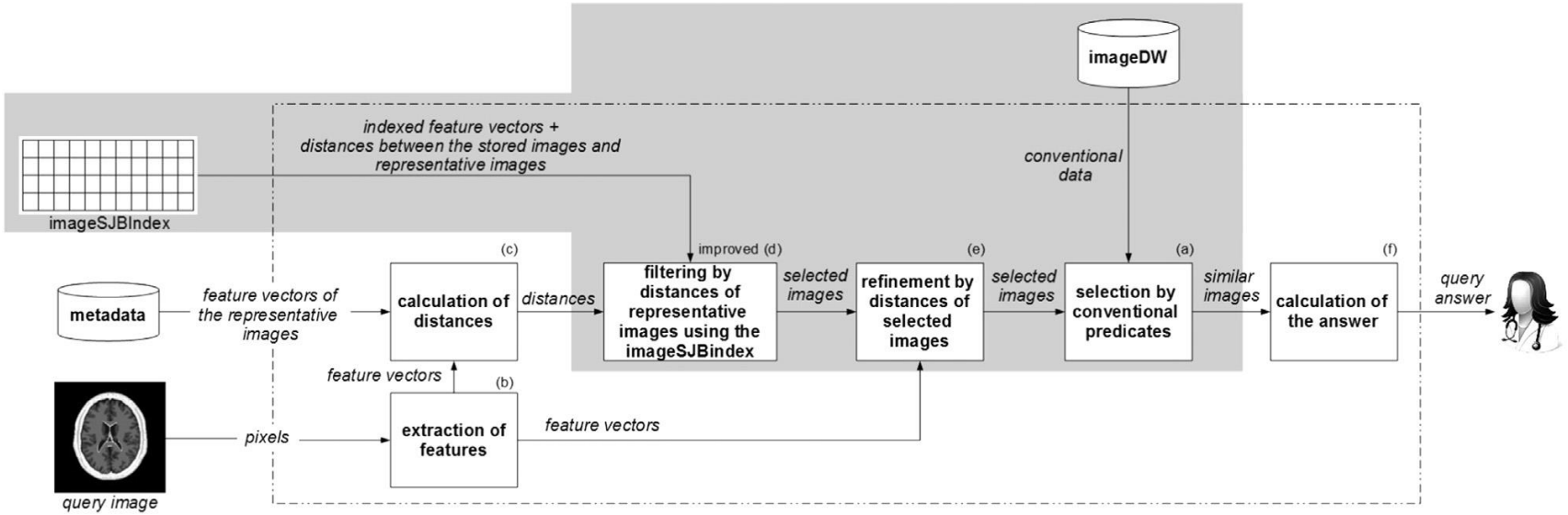
- As imagens são filtradas primeiro pelo pelo predicado de similaridade e, após, pelo predicado convencional
  - *imageSJBindex* → distâncias entre entre as imagens armazenadas e as imagens representativas



Fonte: Teixeira et al. (2015)

splitIndexFV

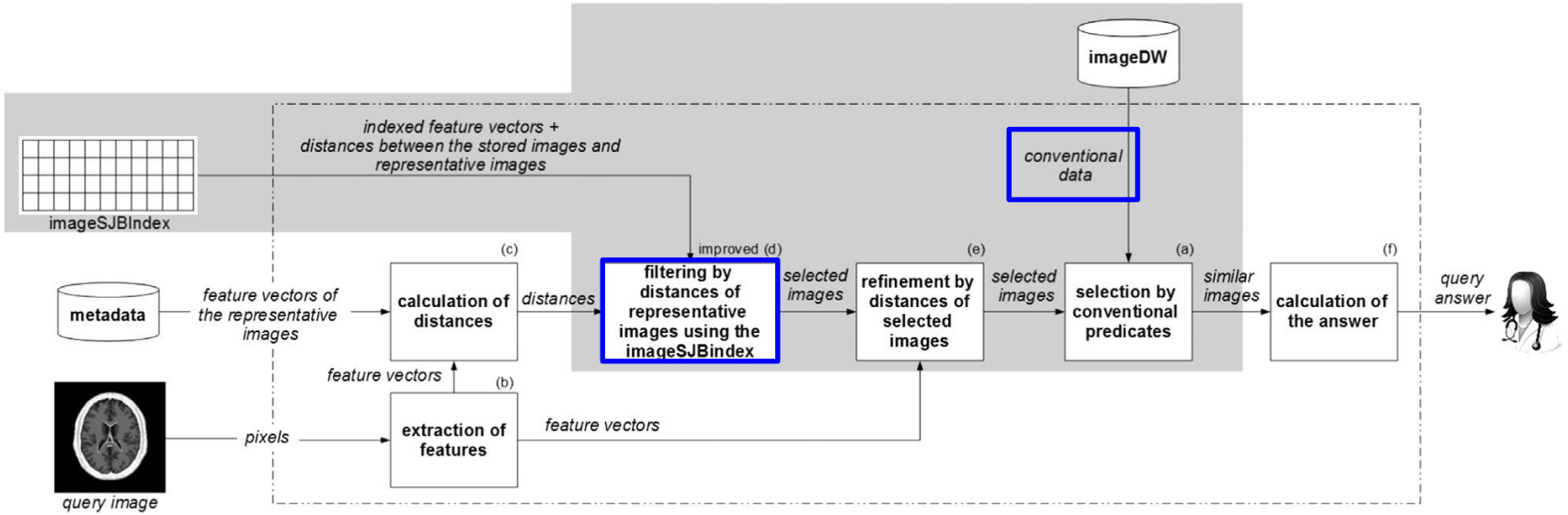
- É a junção das abordagens **dividida (*split*)** com a de **indexar os vetores de características**



Fonte: Teixeira et al. (2015)

splitIndexFV

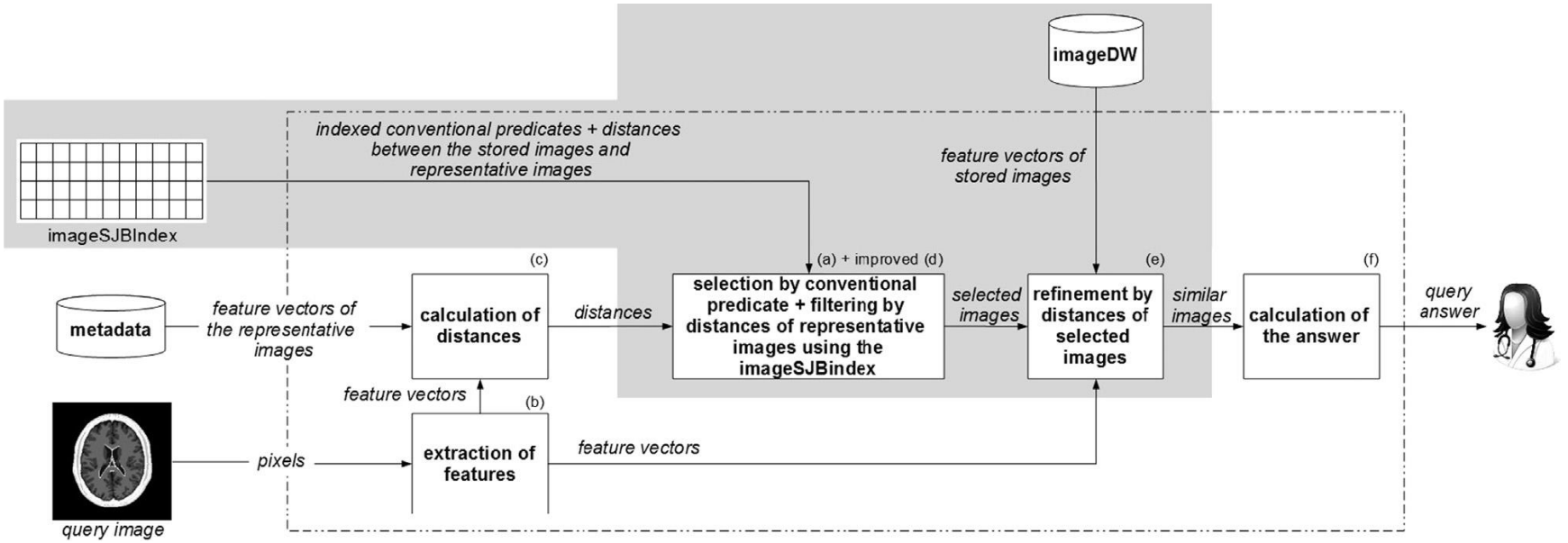
- As imagens são filtradas primeiro pelo pelos predicados de similaridade e, após, pelos predicados convencionais
  - *imageSJBindex* → distâncias entre entre as imagens armazenadas e as imagens representativas e vetores de características das imagens



Fonte: Teixeira et al. (2015)

combinedNotIndexFV

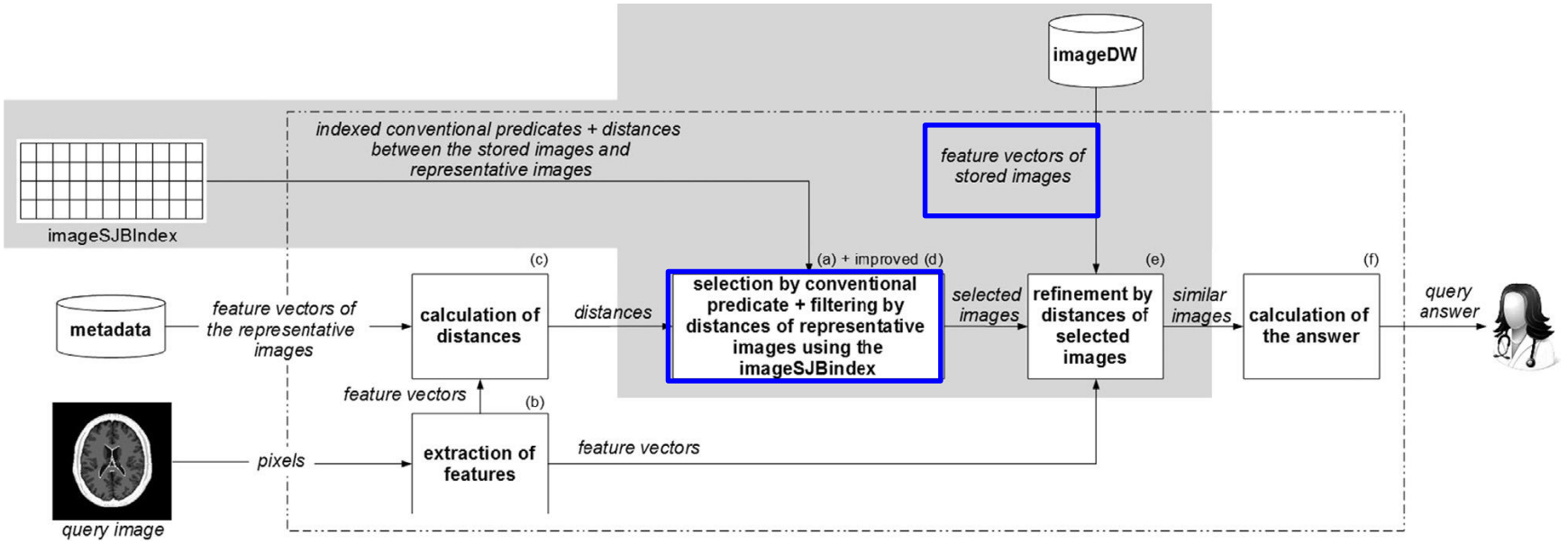
- É a junção das abordagens **combinada (*combined*)** com a de **não indexar os vetores de características**



Fonte: Teixeira et al. (2015)

combinedNotIndexFV

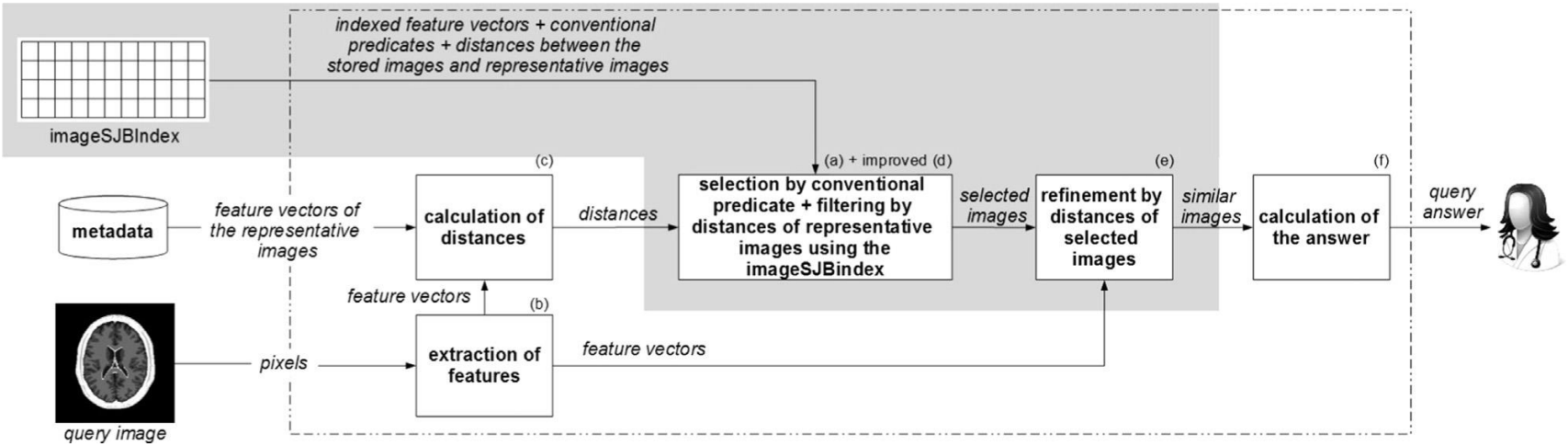
- imagens são filtradas pelos predicados convencionais e os de consulta por similaridade juntos
  - *imageSJBindex* → distâncias entre as imagens armazenadas e as representativas e predicados convencionais



Fonte: Teixeira et al. (2015)

combinedIndexFV

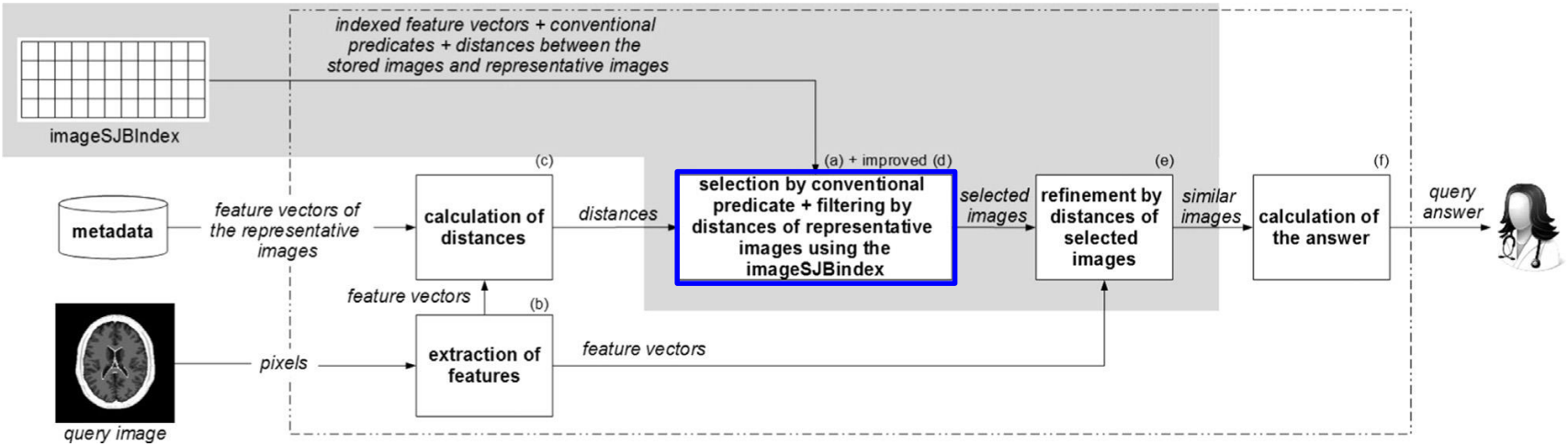
- É a junção das abordagens **combinada (*combined*)** com a de **indexar os vetores de características**



Fonte: Teixeira et al. (2015)

combinedIndexFV

- As imagens são filtradas pelos predicados convencionais e os de consulta por similaridade juntos
  - *imageSJBindex* → todos os dados armazenados no *imageDW*



Fonte: Teixeira et al. (2015)

Comparação das Estratégias de Processamento

Estratégia	Dados Indexados			Acesso ao imageDW	Ordem de Processamento
	Predicado de Busca por Similaridade	Predicado Convencional	Vetor de Características		
splitNotIndexFV	Sim	Não	Não	Sim	(d) melhorado → (a) → (e)
splitIndexFV	Sim	Não	Sim	Sim	(d) melhorado → (e) → (a)
combinedNotIndexFV	Sim	Sim	Não	Sim	(a) + (d) melhorado → (e)
combinedIndexFV	Sim	Sim	Sim	Não	(a) + (d) melhorado → (e)

Fonte: Adaptado de Teixeira et al. (2015)

Resultados

Resultados

- As estratégias *splitNotIndexFV* e *splitIndexFV* são mais adequadas para consultas que contém apenas o predicado de similaridade
  - *splitNotIndexFV* é adequada para vetores de características com altas dimensionalidades
  - *splitIndexFV* é adequada para vetores de características com baixa dimensionalidade

Resultados

- Para consultas que envolvem também predicados convencionais, as estratégias *combinedNotIndexFV* e *combinedIndexFV* são as mais adequadas
  - *combinedNotIndexFV* apresentou melhor desempenhos para vetores de de características com alta dimensionalidade
  - *combinedIndexFV* apresentou melhor desempenhos para vetores de de características com baixa dimensionalidade
  - Apresentaram melhor desempenho quando o predicado convencional é **menos seletivo**

Resultados

- A estratégia básica de processamento se mostrou mais adequada para para consultas envolvendo predicados convencionais e de imagem, quando o predicado convencional é mais seletivo

Referências

- Teixeira, J. W., Annibal, L. P., Felipe, J. C., Ciferri R. R., and Ciferri, C. D. de A. A similarity-based data warehousing environment for medical images. *Comput. Biol. Med.* 66, C (November 2015), 190-208.