Buscas por similaridade com diversificação de resultados apoiadas por índices métricos*

Daniel L. Jasbick[1], **Daniel de Oliveira**[1] **e Marcos V. N. Bedo**[1]

[1]Programa de Pós-Graduação em Computação do Instituto de Computação
Universidade Federal Fluminense

danieljasbick@id.uff.br {marcosbedo, danielcmo}@ic.uff.br

Abstract. *Result diversification enriches k-Nearest Neighbor (k-NN) queries by retrieving objects dissimilar among themselves. Our first contribution, the diversity browsing routine, extends the incremental k-NN distance browsing routing, ensuring result diversification through Influence criteria. Our second contribution leverages diversity browsing to address the search of high-dimensional datasets from the perspective of Local Intrinsic Dimensionality (LID). Our experimental investigation revealed (i) result diversification is constrained by both LID and Influence-based partition principle, (ii) diversity browsing produces query-based manifolds with lower distance concentration regarding the original dataset, (iii) tuned indexes speed up diversity browsing in low/medium LID folds, and (iv) the performance gap between k-NN with and without diversification decreases with LID, with the latter offering richer results.*

Resumo. *A diversificação de resultados enriquece consultas k-NN ao recuperar objetos diferentes entre si. Nossa primeira contribuição, o algoritmo diversity browsing, estende a rotina k-NN incremental distance browsing, garantindo a diversificação de resultados por meio de critérios de Influência. Nossa segunda contribuição analisa o método diversity browsing em espaços de alta dimensionalidade da perspectiva da Dimensionalidade Intrínseca Local (LID). Uma extensa investigação revelou que (i) a diversificação é limitada tanto pela LID quanto pelo particionamento por Influência, (ii) o diversity browsing gera espaços amostrais por consulta com menor concentração de distâncias em comparação aos dados originais, (iii) índices ajustados aceleram o diversity browsing em trechos de baixas/médias LID, e (iv) a diferença de desempenho entre k-NN com e sem diversificação de resultados diminui com a LID.*

1. Introdução

A modelagem por Espaço Métricos oferece uma abordagem eficiente para a comparação de objetos com uma função de distância [Hetland 2009]. A consulta por similaridade aos k-vizinhos mais próximos (k-NN) recupera os k elementos cujas distâncias para um objeto de referência são as menores e é eficientemente executada em Espaços Métricos indexados [Hjaltason and Samet 2003]. Entretanto, buscas k-NN podem gerar resultados semanticamente redundantes (*i.e.*, recuperar objetos que são muito próximos entre si),

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Os autores também agradecem às instituições CNPq (311898/2021-1), FAPERJ (E-26/202.806/2019 - 247357) e FAPESP pelo apoio financeiro.

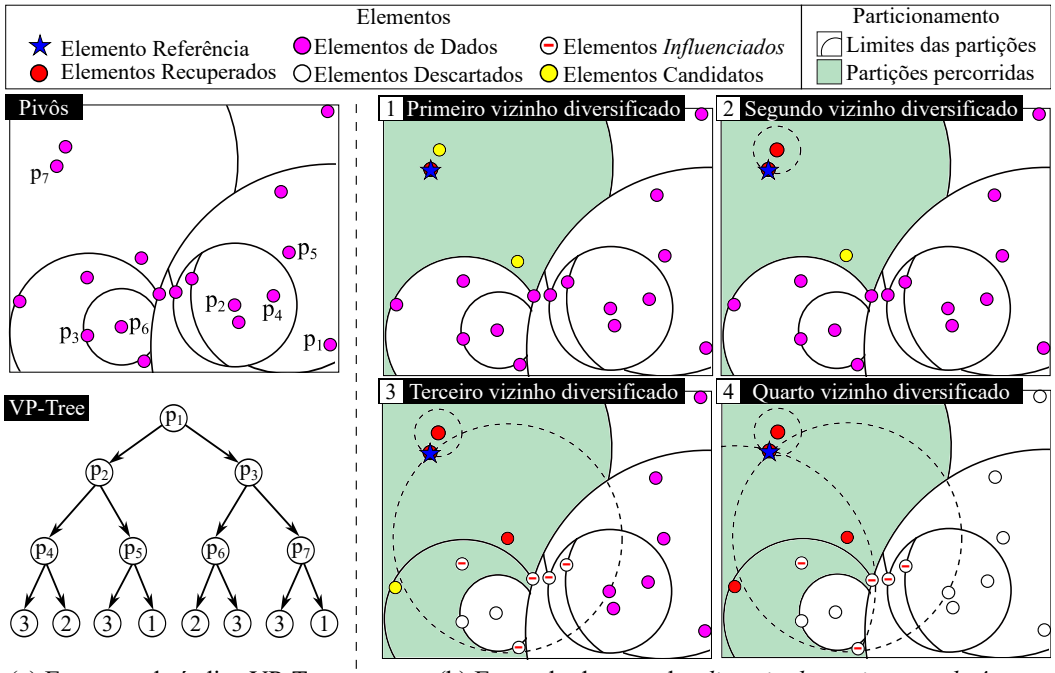
especialmente em espaços de alta dimensionalidade. Neste caso, devido ao fenômeno da concentração de distâncias, à medida que a dimensionalidade dos dados aumenta, a distribuição das distâncias entre pares de pontos tender a convergir para um intervalo estreito de valores, tornando mais difícil distingui-los. Uma alternativa de caráter exploratório à consulta k-NN é a consulta k-NN com diversificação de resultados (dk-NN), que combina o critério de proximidade com critérios de diversidade, para retornar elementos ao mesmo tempo *similares* ao objeto de referência e distintos entre si [Drosou et al. 2017].

A adição de critérios de diversidade, no entanto, aumenta o custo computacional das consultas. Esse custo pode ser reduzido com otimizações nas rotinas de busca em espaços indexados ou nos próprios índices [Kucuktunc and Ferhatosmanoglu 2011]. Uma forma eficiente de integrar critérios de diversidade diretamente na estrutura de busca é agrupando elementos similares e definindo objetos representativos para cada partição. Rotinas baseadas em cobertura [Drosou et al. 2017, Lopes et al. 2021] são boas candidatas para estas de estratégias de *indexar-e-consultar*, pois podem utilizar como critério de diversificação regiões dinâmicas de *Influência* que são descartadas durante a busca [Santos et al. 2013]. Adaptar estes critérios de forma eficiente para espaços puramente métricos, no entanto, ainda era um desafio em aberto na literatura. Assim, a primeira contribuição deste trabalho foi o projeto de um novo algoritmo de busca k-NN com diversificação de resultados apoiado por índices métricos: o método *diversity browsing* [Jasbick et al. 2020]. Este algoritmo estende a rotina de busca k-NN incremental *distance browsing* para consultas dk-NN, combinando os critérios de proximidade e *Influência* para criar *Conjuntos de Influência* que são filtrados em espaços particionados por índices métricos. A implementação[1] do *diversity browsing* foi validada experimentalmente em índices *Vantage-Point Trees* (VP-Trees), se mostrando quantitativamente mais eficiente na execução de consultas dk-NN do que outras abordagens da literatura.

Para avaliar qualitativamente o comportamento do método *diversity browsing*, a rotina foi examinada em condições limite de espaços de alta dimensionalidade, explicitamente identificando objetos em regiões de baixa e alta concentração de distâncias. Uma forma de caracterizar estes objetos é com a medida da *Dimensionalidade Intrínseca Local* (LID), que estima a dificuldade em se distinguir objetos próximos dentro de uma vizinhança para um objeto de referência [Amsaleg et al. 2019]. De forma objetiva, Aumueller e Ceccarello (2021) propõem utilizar a LID como uma estratificação de objetos. Inspirados por este trabalho, avaliamos o comportamento do *diversity browsing* em espaços de alta dimensionalidade, estratificando os conjuntos de dados em *quartis* de LID e examinando o desempenho das buscas sobre nos estratos [Jasbick et al. 2023]. Nessa avaliação, foram identificados *(i)* uma correlação positiva entre o número de vizinhos diversificados e a LID do espaço de busca, *(ii)* que o *diversity browsing* retorna elementos menos concentrados (em termos de distância) do que o os estratos sobre os quais as consultas foram realizadas, *(iii)* que o ajuste de parâmetros de índices métricos ainda é importante em conjuntos reais de alta dimensionalidade e *(iv)* que em casos extremos de LID, o custo do *diversity browsing* é próximo ao do *distance browsing*.

Este artigo está organizado em quatro seções, além da Introdução. A Seção 2 aborda a rotina *diversity browsing*. A Seção 3 apresenta a avaliação experimental. A Seção 4 discute trabalhos relacionados e, por fim, a Seção 5 conclui o artigo.

[1]Disponível em `github.com/UFFeScience/diversity_browsing`



(a) Estrutura do índice VP-Tree (b) Exemplo de consulta *diversity browsing* para k=4

Figura 1. Exemplo de execução do *diversity browsing* sobre um conjunto de dados (parâmetros $k = 4$, distância L_2) indexado por um índice VP-Tree.

2. A rotina *diversity browsing*

Resolver uma consulta dk-NN requer ordenar os objetos do conjunto consultado por suas distâncias ao objeto de referência da busca. Durante esta ordenação, se um objeto está no *Conjunto de Influência* de qualquer outro objeto previamente admitido no conjunto resposta, então ele deve ser descartado da resposta como *Influenciado*. O primeiro objeto da lista é o vizinho inicial e os próximos $k - 1$ objetos não filtrados são os vizinhos diversificados restantes. Portanto, se o conjunto de dados estiver particionado, as partições devem ser ordenadas *antes* dos objetos para garantir que apenas elementos em regiões que não estejam mais afastadas do que o k-ésimo vizinho sejam comparados por distância.

O algoritmo *diversity browsing* proposto durante este trabalho utiliza duas filas de prioridade para ordenar *(i)* as partições do índice e *(ii)* os elementos comparados por distância [Jasbick et al. 2023]. O algoritmo inicia com a avaliação do nó raiz do índice e prossegue iterativamente até encontrar k objetos ou avaliar todas as partições. A cada iteração, se a partição no topo da primeira fila não for *Influenciada* e estiver mais próxima do objeto de consulta do que o elemento no topo da segunda fila, então a partição é removida e seus elementos são inseridos na segunda fila[2]; caso contrário, examina-se o primeiro elemento na segunda fila e, apenas caso ele esteja fora da *Influência* dos demais elementos no conjunto resposta, ele é retornado como o seguinte vizinho diversificado.

A Figura 1(a) apresenta um conjunto bidimensional fictício indexado por uma VP-Tree com seleção aleatória de pivôs e restrição *(i)* de um elemento mínimo por nó-folha e *(ii)* balanceamento obrigatório. A Figura 1(b) apresenta uma consulta dk-NN sobre

[2]No caso de um índice métrico hierárquico, se a partição corresponder a um nó não-folha, então a partição é removida e as suas filhas são inseridas na primeira fila.

esse conjunto para um elemento de referência e parâmetros $k = 4$ e função de distância L_2. Inicialmente, o nó-raiz p_1 é inserido na fila de partições, sendo substituído por seus filhos p_2 e p_3 ordenados por limites de distância mínima e máxima[3]. Posteriormente, p_3 é colocado no topo da fila e substituído por seus filhos p_6 e p_7, com p_7 posicionado ao topo. Seguindo a mesma lógica, p_7 é removido e substituído por dir_{p_7} e esq_{p_7}, estando dir_{p_7} no topo. Sendo um nó-folha, a abertura de dir_{p_7} resulta na inserção de seus elementos na fila de candidatos, deixando três objetos como candidatos ao conjunto resposta. O primeiro vizinho diversificado é recuperado (o próprio pivô p_7), bem como o segundo vizinho na sequência, pois também está mais próximo do elemento de consulta do que qualquer outro objeto e não está *Influenciado* por p_7, como nas Figuras 1(b)[1–2].

A sequência de operações continua com a abertura da partição p_6, já que seus limites indicam que pode conter elementos mais próximos da referência do que o topo da fila de candidatos. Seus filhos, dir_{p_6} e esq_{p_6}, entram na fila de partições com esq_{p_6} no topo. Aqui, os elementos de esq_{p_6} são inseridos na fila de candidatos, mas a prioridade de distância é mantida pelo candidato da partição dir_{p_7}, que é recuperado como segundo vizinho diversificado, como ilustrado na Figura 1(b)[3]. Com a região de *Influência* do segundo objeto recuperado, diversos elementos são eliminados da busca, bem como a partição dir_{p_6}, completamente coberta pela área de *Influência*. Portanto, o nó-direito inteiro dir_{p_6} é descartado da busca. Finalmente, o elemento no topo da fila é recuperado como o terceiro vizinho diversificado, pois não está *Influenciado* e nenhum elemento ou partição está mais próximo da referência do que ele, como ilustrado na Figura 1(b)[4].

3. Avaliações Experimentais

3.1. Conjuntos de Dados e Infraestrutura

Todos os métodos avaliados foram implementados em um único *framework* usando a linguagem Java com JDK 13. Os experimentos foram conduzidos no *cluster* ANOTi[4], composto por dois nós. Cada nó está equipado com 48 *cores* AMD Opteron 6320, 96GB de memória compartilhada, um disco SATA de 1TB e o sistema operacional QLustar. A Tabela 1 resume todos os conjuntos de dados usados nos experimentos.

Tabela 1. Conjuntos de dados usados nas avaliações experimentais.

Nome	Tamanho	Amostra 1	Amostra 2	Dim.	F. Dist.	ID	Avaliação
SINT10	$1 \cdot 10^6$	$70 \cdot 10^3$	$7 \cdot 10^3$	10	L_1	10	Quantitativa
MNIST	$70 \cdot 10^3$	$70 \cdot 10^3$	$7 \cdot 10^3$	784	L_2	12	Quantitativa
YAHOO	$2 \cdot 10^6$	$70 \cdot 10^3$	$7 \cdot 10^3$	400	L_2	08	Quantitativa
COPHIR	$10 \cdot 10^6$	$70 \cdot 10^3$	$7 \cdot 10^3$	282	L_1	15	Quantitativa
MNIST	$70 \cdot 10^3$	$70 \cdot 10^3$	N/A	784	L_2	12	Qualitativa
SIFT	$1 \cdot 10^6$	$140 \cdot 10^3$	N/A	128	L_2	14	Qualitativa
COPHIR	$10 \cdot 10^6$	$280 \cdot 10^3$	N/A	282	L_2	15	Qualitativa

[3]Mais detalhes sobre as distâncias mínimas e máximas em uma VP-Tree em [Jasbick et al. 2020].
[4]Cluster Laboratório ANOTi INFES/UFF: anotilab.com

3.2. Avaliação quantitativa

A avaliação quantitativa foi realizada sobre um conjunto de dados sintético (denominado SINT10[5]) e três conjuntos de dados reais (MNIST[6], YAHOO[7], e COPHIR[8]). A Tabela 1 caracteriza esses conjuntos de dados em termos de sua cardinalidade original (Tamanho), cardinalidade da amostra para as consultas com o *diversity browsing* (Amostra 1), dimensionalidade (Dim.), Dimensionalidade Intrínseca (ID, calculado como em [Chávez et al. 2001]) e tamanho das amostras necessárias para a aplicação dos métodos competidores baseados em diversificação de resultados por novidade (Amostra 2). A Amostra 2 é essencial nessa avaliação devido à complexidade computacional dos métodos GMC e GNE que, sem amostragem, não executariam em um período de tempo aceitável (semanas). Para simplificar o processo, a mesma função de distância foi utilizada para avaliar tanto a *similaridade* quanto a *diversidade* na parametrização dos métodos GMC e GNE. Quanto à avaliação qualitativa, foram utilizados os conjuntos de dados MNIST[6], COPHIR[8] e SIFT[9], todos imersos em espaços de alta dimensionalidade.

Medidas de desempenho. A eficiência do *diversity browsing* foi analisada experimentalmente através de três métricas: tempo de execução das consultas, número de cálculos de distância realizados e quantidade de nós visitados. O tempo de execução das consultas oferece uma medida do custo associado a cada consulta, embora esteja relacionado à infraestrutura de *hardware* utilizada. Ao quantificar o número de cálculos de distância, podemos estimar o custo computacional das consultas de uma maneira independente da infraestrutura. Adicionalmente, o total de nós visitados durante a busca possibilita a avaliação da eficiência do *diversity browsing* na exploração de partições de VP-Trees e, além disso, auxilia no entendimento de quão eficientes são diferentes parametrizações dos índices na segmentação dos conjuntos de dados.

Ajuste dos índices. São explorados nessa análise os critérios de seleção do pivô e o suporte de excesso (*overflow*) nos nós-folha de VP-Trees. Os critérios de seleção de pivôs comparados foram: escolha aleatória com distribuição uniforme iid. (RND); de fecho-convexo (CVX); ou variância máxima (MAX). Outro parâmetro avaliado foi a opção pelo balanceamento (BAL) ou não balanceamento (UBAL) da árvore, de forma que a opção BAL deve suportar o *overflow* de elementos nos nós-folha.

Eficiência das consultas. Os resultados revelam que o *diversity browsing* com VP-Trees VP_MAX_BAL foi eficiente superando outras configurações. A Figura 2 ilustra os resultados deste experimento em termos de tempo de execução. Embora o impacto do suporte a *overflow* tenha sido menos significativo em comparação com os critérios de seleção do pivô (*i.e.*, o método de seleção de pivôs dominou o desempenho), uma diferença sutil foi observada entre as estruturas de variância máxima. Esta descoberta sugere que a escolha da parametrização do VP-Tree é crucial na otimização de buscas dk-NN. Portanto, o uso de VP-Trees VP_MAX_BAL é aconselhável para uma execução eficiente.

Comparação de desempenho. Nessa avaliação o *diversity browsing* apresentou desempenho superior ao método baseline (BRID$_k$ [Santos et al. 2013, Santos et al. 2018])

[5]Dimensões *iid.* no intervalo [0,1]. Disponível em sites.labic.icmc.usp.br/mldatagen
[6]Dígitos manuscritos. Disponível em yann.lecun.com/exdb/mnist
[7]Atributos de imagens do Yahoo®. Disponível em webscope.sandbox.yahoo.com
[8]Atributos de cor de fotografias. Disponível em cophir.isti.cnr.it
[9]Atributos de baixo nível de imagens SIFT. Disponível em corpus-texmex.irisa.fr

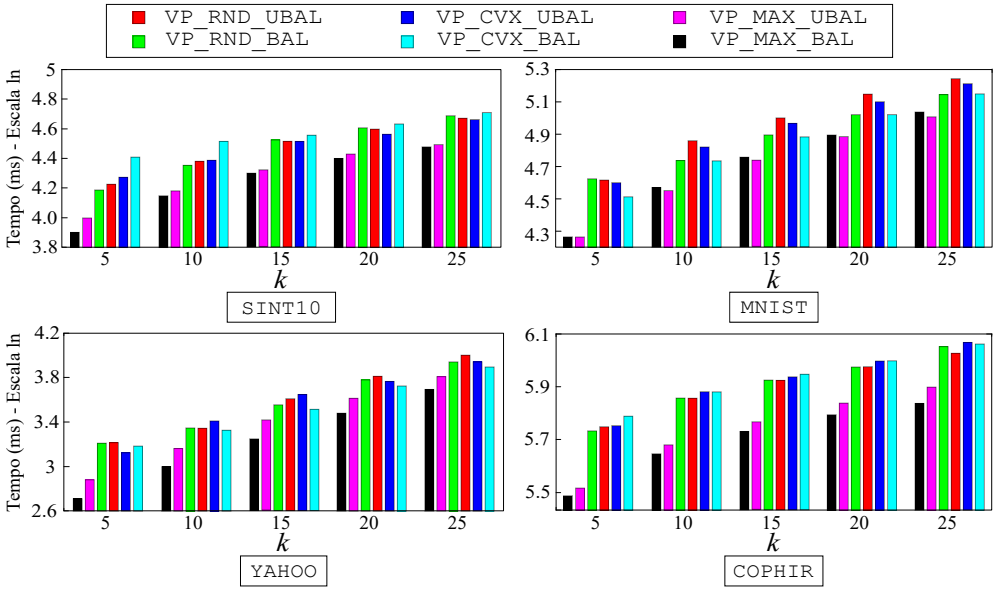


Figura 2. Média de execução do *diversity browsing* em diferentes VP-Trees.

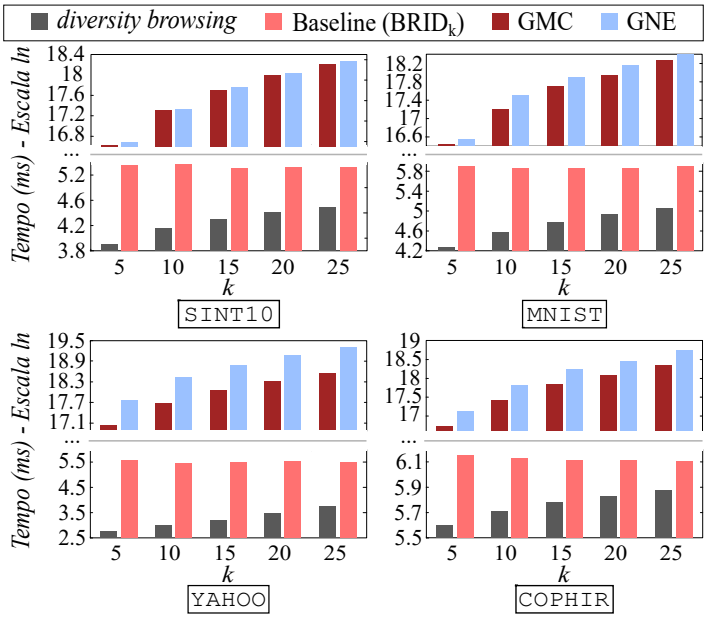


Figura 3. Média de tempo de consultas: *diversity browsing vs.* competidores.

e competidores baseados em novidade *Greedy Marginal Contribution* GMC e *Greedy Randomized with Neighborhood Expansion* (GNE) [Vieira et al. 2011] em termos de cálculos de distância e tempo de consulta. Mais especificamente, exigiu menos cálculos de distância por um fator de até aproximadamente $15\times$, dependendo do conjunto de dados. A abordagem proposta também foi pelo menos $87{,}51\%$ mais rápida que o BRID$_k$. A vantagem de desempenho sobre GMC e GNE foi ainda mais notável, com o *diversity browsing* sendo, no mínimo, em torno de 5 ordens de magnitude mais rápido.

Custos de construção. Em termos de custos de construção, os resultados indicam que VP-Trees com pivôs de variância máxima geram melhor desempenho para consul-

tas com o *diversity browsing*, mas também causam um aumento significativo no custo de construção. No entanto, mesmo com o custo mais elevado, a rotina com melhor construção se mostrou valiosa após um certo número de consultas realizadas. Por exemplo, considerando tanto o tempo de consulta quanto o custo de construção do índice, o *diversity browsing* se tornou mais eficiente que o BRID$_k$ após a execução de algumas milhares de consultas (entre 1061 e 2379), o que representa uma pequena porcentagem (em torno de $2{,}74\%$) do total de elementos indexados. Mesmo ao usar VP-Trees menos custosas (pivôs aleatórios), no entanto, a rotina conseguiu manter uma vantagem de desempenho, sendo pelo menos 14% mais rápido que o BRID$_k$. Isso destaca a robustez e versatilidade do *diversity browsing* na comparação com os métodos que existiam na literatura antes da proposta deste trabalho [Jasbick et al. 2020].

3.3. Avaliação qualitativa

Com o objetivo de caracterizar o comportamento da rotina *diversity browsing* proposta foi realizada uma série de avaliações experimentais em cenários limites para a proposta: conjuntos de dados imersos em espaços de alta dimensionalidade. Todas as avaliações foram realizadas usando uma estratégia de *holdout*, com 10% dos dados selecionados aleatoriamente como objetos de consulta e os 90% restantes utilizados para indexação e consulta. As rotinas de busca *distance* (k-NN) e *diversity browsing* (dk-NN) foram comparadas sobre subconjuntos estratificados em quatro categorias com base nos quartis das distribuições de Dimensionalidade Intrínseca Local (LID), calculadas pela aproximação em [Amsaleg et al. 2019] e cada uma representando faixas de concentração de distância.

O primeiro quartil (1QT) é compostos pelos elementos de menor LID, considerados "mais fáceis" (menos concentrados), enquanto o segundo quartil (2QT) é composto pelos objetos entre o 1QT e a mediana da distribuição de LID. O terceiro quartil (3QT) é formado pelos elementos entre a mediana e o último quartil, enquanto que o quarto quartil (4QT) é formado pelos elementos com LID superior aos do 3QT, considerados "mais difíceis" (mais concentrados). Este método de classificação de elementos "mais fáceis" e

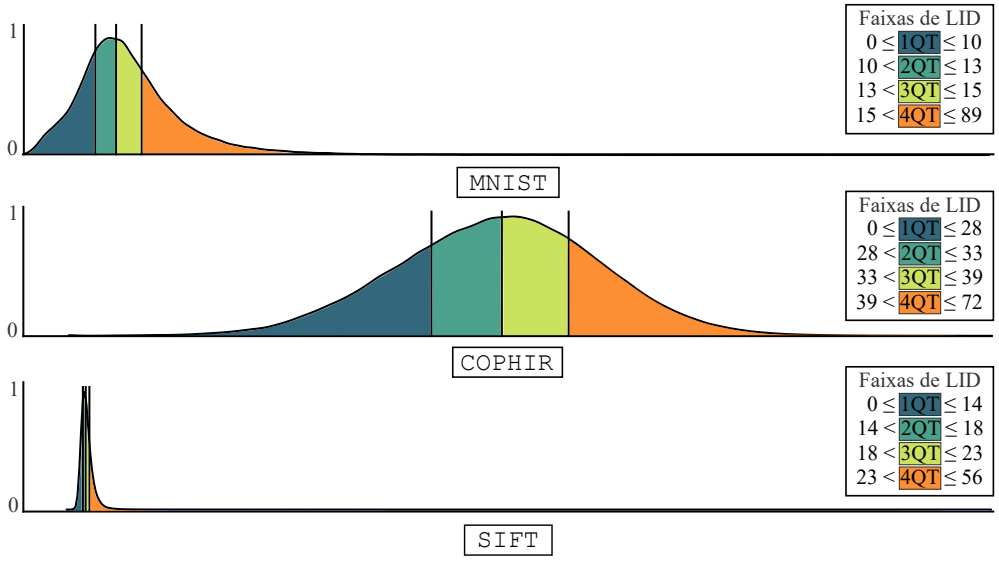


Figura 4. Distribuição de valores de Dimensionalidade Intrínseca Local.

"mais difíceis" em conjuntos separados permite um exame controlado do comportamento de diferentes intervalos de LID, evitando que dados de LID mais baixas afetem os resultados de dados de LID mais altas e vice-versa. A Figura 4 apresenta as distribuições de LID para cada um dos conjuntos de dados examinados, considerando cada elemento como um objeto de consulta e vizinhança $k = 100$ [Amsaleg et al. 2015].

O objetivo desta avaliação estratificada é caracterizar o *diversity browsing* obtendo indícios experimentais que auxiliem a responder as seguintes questões em aberto:

- **(Q1)** A quantidade de vizinhos diversificados recuperáveis está relacionada a LID?
- **(Q2)** Até que ponto os *manifolds* estão concentrados em termos de distância?
- **(Q3)** O ajuste dos índices pode melhorar o desempenho do *diversity browsing* mesmo para conjuntos de alta dimensionalidade?
- **(Q4)** Como o *diversity browsing* se comporta comparado ao *distance browsing* em quartis de LID muito altas ou baixas?

Medidas de qualidade. Para tratar destas questões em aberto, foram tomadas diferentes medidas de qualidade sobre os conjuntos de dados originais, quartis e sub-espaços (*manifolds*). Especificamente, foram mensuradas a quantidade de vizinhos recuperáveis e as medidas de concentração de distâncias de Variância Relativa (RV) [Francois et al. 2007], Contraste Relativo (RC) [Beyer et al. 1999] e Dimensionalidade Intrínseca (ID) [Chávez et al. 2001, Pestov 2013]. A medida RV mede proporção existente entre a variância e a média das distribuições de distância entre pares de objetos do conjunto analisado. Já a medida RC permite a avaliação da diferença entre a distância mínima e máxima presentes nas distribuições, auxiliando na compreensão da extensão da variação dentro do conjunto de dados analisado [Beyer et al. 1999, Francois et al. 2007]. Adicionalmente, incluiu-se na análise a ID, empregada para se mensurar a concentração de distância no conjunto de dados de formal global [Chávez et al. 2001].

Correlação entre LID e diversidade. Nessa avaliação, foram realizadas consultas com o *diversity browsing* para valores de $k \to \infty$, ou seja, fazendo como que o parâmetro k fosse configurado para o limite máximo de número de vizinhos diversificados (no máximo, o tamanho do quartil de dados). Esta configuração permite uma busca exaustiva em todo o espaço disponível, com exceção aos elementos *Influenciados* por vizinhos já recuperados durante a rotina incremental. A Figura 5 mostra a quantidade de de vizinhos recuperados para valores crescentes de LID na forma de box-plots sem poda em valores extremos. Os resultados sugerem uma correlação entre a LID e o número de vizinhos, *i.e.*, LIDs mais

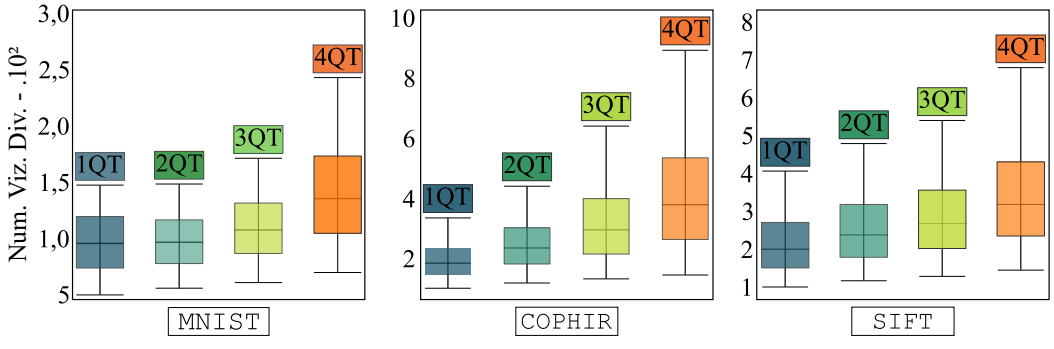


Figura 5. Número de vizinhos obtidos em consultas com $k \to \infty$.

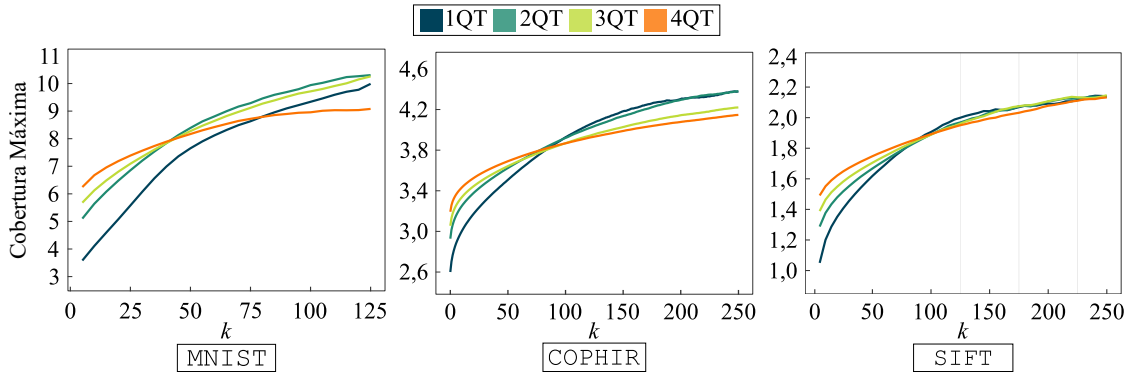


Figura 6. Cobertura Máxima dos k-ésimos Conjuntos de *Influência* para todos os conjuntos de dados e intervalos de LID examinados.

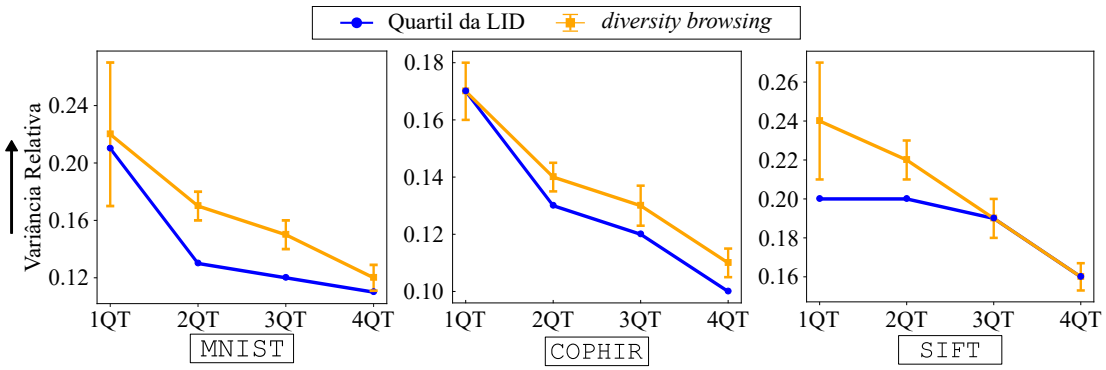


Figura 7. Comparações da Variância Relativa entre quartis estratificados e os *manifolds* obtidos pelo *diversity browsing*.

altas resultaram em mais diversidade do que LIDs mais baixas (Pearson $\texttt{MNIST} \approx 0.31$, $\texttt{SIFT} \approx 0.25$, $\texttt{COPHIR} \approx 0.41$). Além disso, mais vizinhos foram encontrados dentro de quartis com LIDs mais altas em todos os casos examinados. Independentemente da LID, o máximo de vizinhos recuperados pelo *diversity browsing* foi (muito) menor do que o tamanho inicial do quartil, ressaltando a capacidade dos critérios de *Influência* em selecionar os objetos no conjunto resposta. No caso do MNIST, por exemplo, não mais que 265 objetos foram recuperados do 4QT, sendo que cada quartil contém $15{,}750$ elementos.

A Figura 6 mostra a razão para esse comportamento: a distância entre o objeto de consulta e seu k-ésimo vizinho diversificado define uma "Cobertura Máxima" para a qual objetos podem ser descartados como *Influenciados*. Para cada valor de k, a figura ilustra a mediana da Cobertura Máxima atingida pelas consultas com o *diversity browsing* em cada quartil, sendo que houve um aumento da cobertura em relação ao valor de k de forma mais rápida e acentuada para consultas em quartis de menor LID (descartando regiões maiores e, consequentemente, mais elementos) e um aumento mais lento para consultas em quartis de maior LID (descartando regiões menores e, consequentemente, menos elementos. Ou seja, mais diversidade pode ser encontrada em espaços de maior dimensionalidade, já que a *Influência* definida por cada vizinho encontrada é menor.

Medidas de qualidade sobre os *manifolds* gerados pelo *diversity browsing*. A execução de consultas *diversity browsing* com $k \to \infty$ geram subespaços (*manifolds*) orientados ao

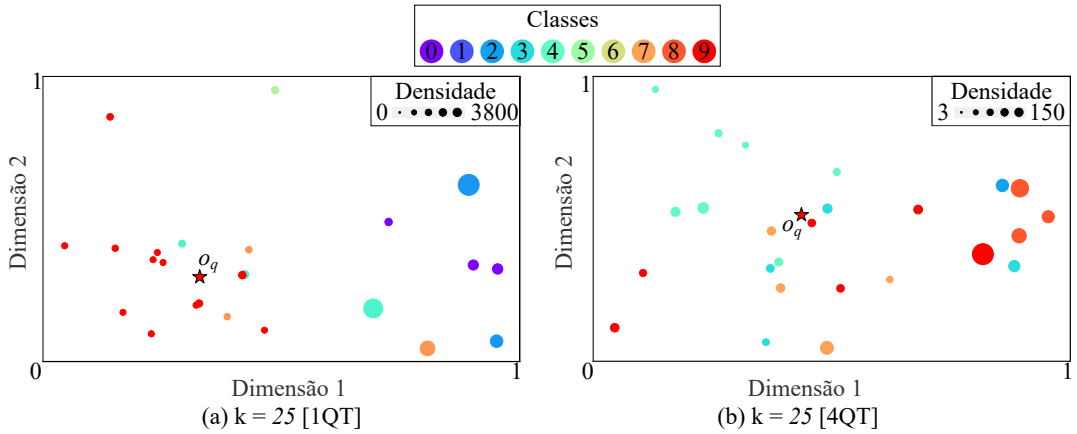


Figura 8. Vizinhança diversificada $k = 25$ para um mesmo meta-atributo sobre o 1QT e 4QT do conjunto de dados `MNIST`.

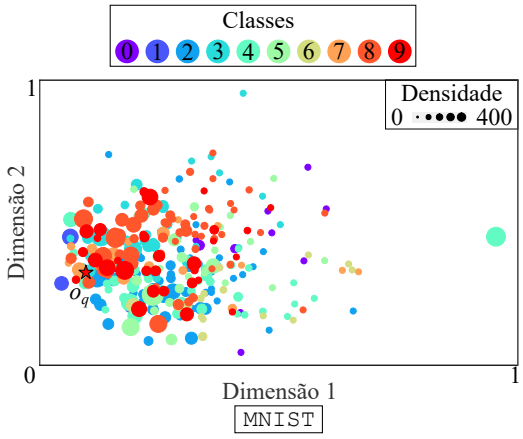


Figura 9. Visualização de uma vizinhança $k \to \infty$ para o 4QT do `MNIST`.

objeto de consulta, onde espera-se que a semântica de objetos "próximos" e "afastados" seja reforçada, *i.e.*, geram distribuições de distância entre elementos menos concentradas que no conjunto original. A Figura 7 quantifica essa observação por quarties de LID considerando a medida de Variância Relativa (quanto maior a VR, menos concentrada está a distribuição de distâncias). Os *manifolds* apresentaram até 20% menor concentração em termos de Variância Relativa (RV) para o conjunto de dados `MNIST` e uma Dimensionalidade Intrínseca (ID) reduzida de até 27% [Jasbick et al. 2023]. Esta tendência se manteve para os conjuntos de dados `SIFT` e `COPHIR`, com concentrações até 17% e 9% menores e IDs menores em até 34% e 12%, respectivamente. A medida de RV sugere uma separação consistente entre os quartis e as distribuições de distâncias dos *manifolds*, que exibiram média inferior e maior variância. Além disso, houve uma tendência às LIDs menores levarem para distribuições de distância mais esparsas.

Ao se comparar medidas de RV dos quartis e dos *manifolds* de cada objeto de consulta, constata-se que a medida de Contraste Relativo (RC) dos conjuntos originais excede a dos quartis, que supera a dos *manifolds*. Isso é explicado pela sensibilidade da medida à magnitude das distâncias máximas e mínimas, que é maior no conjunto original do que nos quartis e menor nos *manifolds*. Contudo, os *manifolds* alcançaram valores de

RC comparáveis àqueles de suas contrapartes em relação a LID, com a variação de RC dos *manifolds* em aproximadamente 63%, comparada à variação de 65% entre o 1QT e 4QT do conjunto de dados `MNIST`.

Uma consequência desses resultados (*diversity browsing* sendo capaz de encontrar *manifolds* menos concentrados para cada objeto de consulta) é que a visualização dos conjuntos resposta tende a fornecer informações interessantes sobre a proximidade dos dados e que a redução de dimensões tende a gerar representações menos sobrepostas. Para ilustrar essa consequência, foi proposta a visualização dos *manifolds* encontrados na forma de um *gráfico de bolhas*, com cada bolha centrada no objeto do conjunto resposta reduzido para duas dimensões e com o raio da bolha sendo proporcional à quantidade de vizinhos *Influenciados* pelo objeto do conjunto resposta. As Figuras 8(a–b) apresentam essa visualização com o apoio de uma redução PCA para duas dimensões para o conjunto de dados `MNIST` considerando as faixas de concentração 1QT e 4QT.

No conjunto de dados analisado, cada elemento é vinculado a um "dígito" como meta-atributo. Para fins de visualização, esses atributos são representados nas bolhas por diferentes cores, com o objeto de consulta definido como o dígito "9". Da visualização, se percebe que os vizinhos que pertencem à classe "9" tendem a estar mais próximos e são mais recorrentes quando comparados a outros conceitos no 1QT. Além disso, há uma maior densidade de bolhas no 1QT, que diminui conforme a LID aumenta, permitindo que um maior número de elementos seja abrangido no 1QT (0–3800) do que no 4QT (3–150). A análise também revela que os meta-atributos "2" e "8" são mais comuns no 1QT, enquanto o "4" e "3" passam a aparecer mais frequentemente à medida que a LID aumenta, com o valor de k constante. Na análise do 1QT, nota-se que os vizinhos mais próximos estão associados ao conceito mais comum, "9", enquanto no 4QT, os vizinhos mais próximos estão mais relacionados a conceitos como "3" "4" e "7". A visualização complementar, apresentada na Figura 9, exibe os *manifolds* que resultam de consultas ao 4QT com $k \to \infty$. Esta representação ressalta a noção de proximidade entre o objeto de consulta e os outros elementos, destacando a menor concentração de distâncias nos *manifolds*. Portanto, conclui-se que, mesmo com o 4QT retornando o maior número de vizinhos dentre os quartis avaliados (conforme discutido na Seção 3.3), a separação das distâncias entre o elemento de consulta e os vizinhos recuperados permanece visível.

Ajuste fino do índice. Em espaços de alta dimensionalidade, a eficiência do índice tende a degradar, eventualmente para uma busca sequencial. Esta avaliação, portanto, tem como

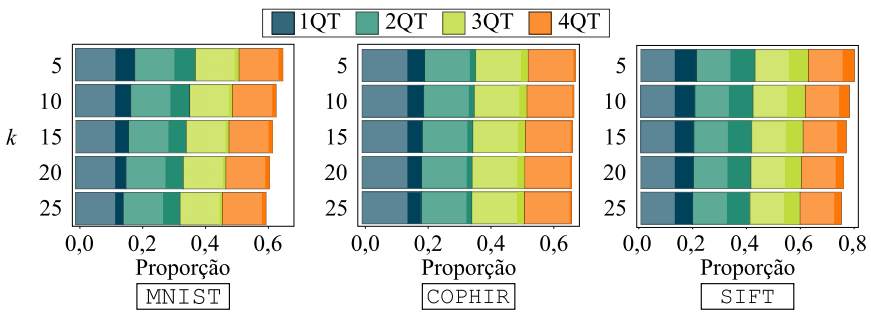


Figura 10. Comparações de desempenho com ajuste de VP-Trees entre métodos de seleção de pivôs de Variância Máxima e Aleatório.

objetivo verificar a aplicabilidade prática do índice em diferentes estratificações de LID. Para isso foi examinado novamente o índice VP-Tree com configuração com pivôs de *Variância Máxima* e aleatórios em um árvore balanceada (`VP_MAX_BAL`, `VP_RND_BAL`), dado que esta configuração requer menos cálculos de distância e tempo de execução do que as outras (ver Figura 2). Os ganhos obtidos pela Variância Máxima ilustrados na Figura 10 (proporção de objetos de consulta para os quais a Variância Máxima superou a escolha aleatória em cada quartil) mostram o ajuste do índice é sim capaz de otimizar a busca. Os maiores ganhos foram observados no 1QT, onde as buscas com a configuração `VP_MAX_BAL` proporcionaram um aumento de eficiência em relação aos 2QT e 3QT. Estes resultados sugerem que a otimização de índices métricos continua sendo um fator importante na otimização de buscas dk-NN em conjuntos de alta dimensionalidade, já que trechos destes conjuntos não são altamente concentrados. Adicionalmente, os resultados indicam que o *diversity browsing* pode ser beneficiado por um eventual particionamento do conjunto por LID, onde para cada pedaço do conjunto de dados uma estrutura de indexação pode ser usada para melhorar o desempenho geral da consulta.

Desempenho de consultas por similaridade e similaridade diversificada. A Figura 11 mostra as diferenças de desempenho entre o *distance* e *diversity browsing* no contexto da execução de buscas no 1QT e 4QT. Os resultados indicam um aumento nas variações de desempenho conforme cresce o número de elementos recuperados, assim como uma sensibilidade do desempenho do *diversity browsing* em relação ao tamanho da vizinhança nas consultas realizadas no 1QT. Essa sensibilidade está ligada ao critério de cobertura, uma vez que o espaço navegado para recuperar os k elementos mais próximos com o *distance browsing* é menor do que o espaço que o *diversity browsing* inspeciona para retornar os k vizinhos diversificados mais próximos. O 4QT revela contrastes mais sutis, com disparidades de desempenho para valores de k acima de 15. A complexidade inerente ao espaço de alta dimensionalidade exige que ambas rotinas inspecionem uma região similar do espaço de busca. Como resultado, o número de cálculos de distância para ambas as rotinas é similar, com o *diversity browsing* sendo, no máximo, $8{,}4\%$ mais custoso do que o *distance browsing*. Consequentemente, nota-se que o *diversity browsing* incorre em custos mais altos em conjuntos de menor dimensionalidade, mas oferece resultados superiores em termos de concentração de distâncias em espaços de alta dimensionalidade, com custos comparáveis aos do *distance browsing*.

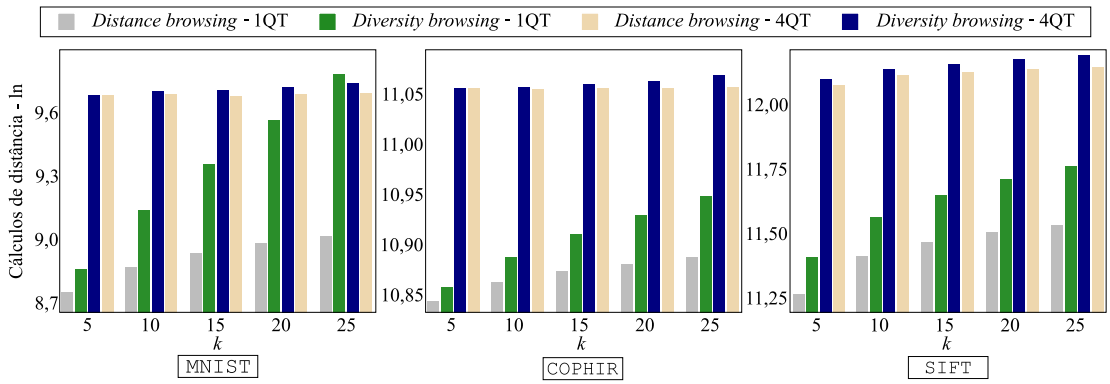


Figura 11. Diferenças de desempenho entre *distance* e *diversity browsing*.

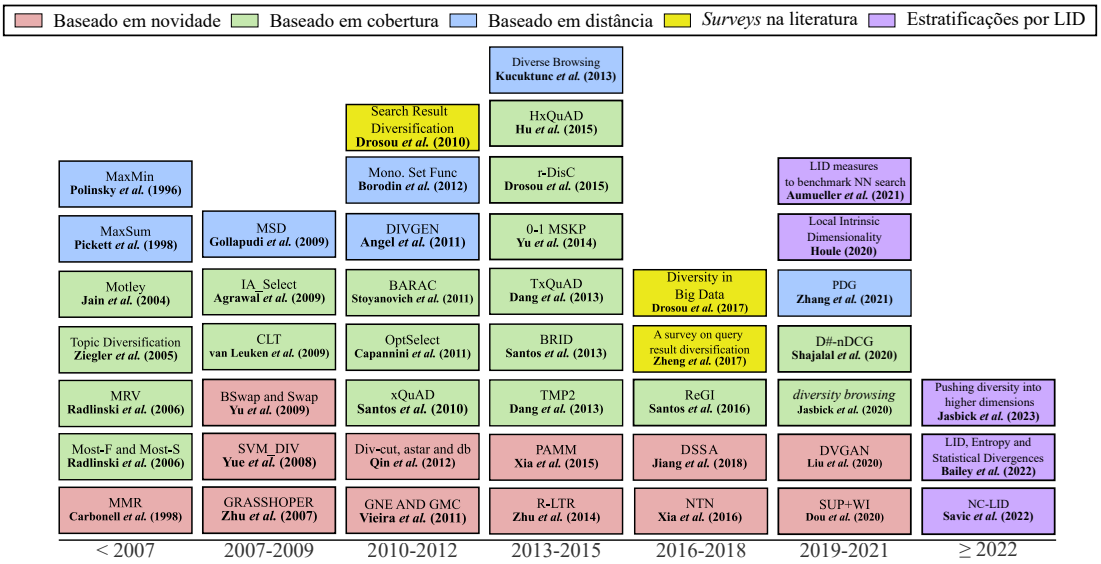


Figura 12. Linha do tempo com os métodos de diversificação de resultados revisados nesse trabalho e trabalhos relacionados.

4. Trabalhos Relacionados

A revisão bibliográfica deste trabalho foi feita com base em mapeamentos da literatura utilizando como base o Google Scholar[10] e a DBLP[11]. Foram utilizados como termos-chave as expressões em inglês *result diversification* e *diversified similarity searching*. Os resultados obtidos passaram por um processo de filtragem, no qual foram eliminados os métodos derivados de áreas não relacionadas à Teoria de Espaços Métricos. Foram selecionados, então, os 45 melhores artigos para serem incluídos na linha do tempo, juntamente com quatro estudos emergentes posteriores ao início deste trabalho. O resultado dessa revisão é apresentada como uma linha do tempo, incluída na Figura 12. De particular importância é o trabalho em Drosou et al. (2017) que apresenta uma taxonomia e um contraste entre algoritmos de busca por similaridade e diversificação de resultados, fornecendo indicações iniciais sobre a correlação entre dimensionalidade e diversidade. Além deste, o tabalho de Aumueller et al (2021) também é de particular importância, ao ser o primeiro a definir uma carga estratificada por LID para caracterizar o comportamento de algoritmos de busca k-NN aproximados. Seguimos as indicações de todos estes trabalhos para a projeto, implementação e avaliação da rotina *diversity browsing*.

5. Conclusões e Trabalhos Futuros

Este trabalho estudou desafios abertos em diversificação de resultados e propôs o algoritmo *diversity browsing*, uma nova abordagem para buscas dk-NN baseada em cobertura incremental eficiente para espaços indexados. A eficácia do método é caracterizada através de uma ampla avaliação experimental quantitativa, onde a rotina é comparada com um método *baseline* (BRID$_k$) e outros dois métodos de diversificação por novidade. Os resultados mostraram que a proposta superou com margem os métodos existentes na literatura. Na sequência foi realizada uma avaliação experimental qualitativa com o objetivo

[10]Disponível em `scholar.google.com.br`
[11]Disponível em `dblp.org`

de caracterizar o comportamento do método *diversity browsing* em cenários extremos de alta dimensionalidade, por meio da estratificação por LID. Os resultados mostraram um correlação entre o estrato LID e a capacidade do *diversity browsing* de encontrar vizinhos diversificados e também que o conjunto resposta encontrado pelo algoritmo proposto pode ser visto como um *manifold* orientado ao objeto de consulta onde as distâncias estão menos concentradas do que no conjunto original e no estrato da LID. Resta ver o potencial desse resultado para visualização de dados e como método de amostra, capaz de gerar amostras *justas* no sentido da distribuição distâncias possuir variância aceitável e discriminativa. Os resultados também mostraram que a realização de buscas indexadas é vantajosas mesmo em espaços de alta dimensionalidade e que o custo de adicionar diversificação de resultados é relativo na comparação com buscas k-NN. Resta investigar a adoção de diferentes índices/métodos de acesso com aderência ao extrato da LID com potencial para melhorar ainda mais o desempenho de buscas dk-NN.

Referências

Amsaleg, L., Chelly, O., Furon, T., Girard, S., Houle, M., Kawarabayashi, K., and Nett, M. (2015). Estimating local intrinsic dimensionality. *Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD)*, pages 29–38. ACM.

Amsaleg, L., Chelly, O., Houle, M., Kawarabayashi, K.-I., Radovanović, M., and Treeratanajaru, W. (2019). Intrinsic dimensionality estimation within tight localities. *International Conference on Data Mining (SDM)*, pages 181–189. SIAM.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? *International Conference on Database Theory (ICDT)*, pages 217–235. Springer.

Chávez, E., Navarro, G., Baeza-Yates, R., and Marroquín, J. (2001). Searching in metric spaces. *Computing Surveys*, volume 33, pages 273–321. ACM.

Drosou, M., Jagadish, H., Pitoura, E., and Stoyanovich, J. (2017). Diversity in big data: A review. *Big data*, 5(2):73–84.

Francois, D., Wertz, V., and Verleysen, M. (2007). The concentration of fractional distances. *Transactions on Knowledge and Data Engineering (TKDE)*, volume 19, pages 873–886. IEEE.

Hetland, M. L. (2009). The basic principles of metric indexing. *Swarm intelligence for multi-objective problems in data mining*, pages 199–232. Springer.

Hjaltason, G. and Samet, H. (2003). Index-driven similarity search in metric spaces. *Transactions on Database Systems (TODS)*, 28(4):517–580.

Jasbick, D., Santos, L., Azevedo-Marques, P. M., Traina, A. J., de Oliveira, D., and Bedo, M. (2023). Pushing diversity into higher dimensions: The lid effect on diversified similarity searching. *Information Systems*, 114:102166.

Jasbick, D., Santos, L., de Oliveira, D., and Bedo, M. (2020). Some branches may bear rotten fruits: Diversity browsing vp-trees. *Similarity Search and Applications (SISAP)*, pages 140–154. Springer.

Kucuktunc, O. and Ferhatosmanoglu, H. (2011). λ-diverse nearest neighbors browsing for multidimensional data. *Transactions on Knowledge and Data Engineering (TKDE)*, volume 25, pages 481–493. IEEE.

Lopes, C., Santos, L., Jasbick, D., de Oliveira, D., and Bedo, M. (2021). An empirical assessment of quality metrics for diversified similarity searching. *Journal of Information and Data Management (JIDM)*, 12(3).

Pestov, V. (2013). Lower bounds on performance of metric tree indexing schemes for exact similarity search in high dimensions. *Algorithmica*, 66(2):310–328.

Santos, L., Blanco, G., Oliveira, D., Traina, A., Traina Jr, C., and Bedo, M. (2018). Exploring Diversified Similarity with Kundaha. *Conference on Information and Knowledge Management (CIKM)*, pages 1903–1906. ACM.

Santos, L., Oliveira, W., Ferreira, M., Traina, A., and Traina Jr, C. (2013). Parameter-free and domain-independent similarity search with diversity. *International Conference on Scientific and Statistical Database Management (SSDBM)*, pages 1–12.

Vieira, M., Razente, H., Barioni, M., Hadjieleftheriou, M., Srivastava, D., Traina Jr., C., and Tsotras, V. (2011). On query result diversification. *International Conference on Data Engineering (ICDE)*, pages 1163–1174. IEEE.