Levantamento de Conjuntos de Dados e Construção de Bases de Características para Avaliação de Técnicas de Busca por Similaridade

Matheus Augusto Leme Matiazzo[1], **Daniel dos Santos Kaster**[1]

[1]Departamento de Computao – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – CEP 86057-970 – Londrina – PR – Brasil

mathmatiazzo@gmail.com, dskaster@uel.br

Abstract. *The last decades required a great development of similarity research methods, this was due to the large production of complex data. Complex data is data rich in content, such as images, audios and videos. Several indexing structures were also studied so that we could improve this query, the one that stood out so far was the proximity graphs. The main objective of this work is to form a dataset that is able to show the diversity of the sets used in the area of data retrieval by similarity, verifying their characteristics and comparing them with the characteristics extracted from other datasets. Also, demonstrate the method that was used to arrive at this final set. With this set assembled and analyzed, we seek to show whether it is possible to improve or perhaps enrich the choice of data sets used in similarity retrieval articles, noting whether there is any trend in this choice, such as the most popular sets. For this, first a survey of complex data sets used by researchers in the field will be carried out, after the survey we will analyze the properties of the data sets themselves. To the best of our knowledge so far in the literature, this analysis and survey of datasets taking into account the properties of the datasets has not yet been done. To the best of our knowledge so far in the literature, this analysis and survey of datasets taking into account the properties of the datasets has not yet been done. As a final result, it is expected to show how the data sets behave, providing the set already extracted and analyzed together with the method used for researchers in the field to analyze their own data sets*

Resumo. *As últimas décadas exigiram um grande desenvolvimento dos métodos de pesquisa por similaridade, isso se deu devido a grande produção de dados complexos. Dados complexos são dados ricos em conteúdo, como por exemplo imagens, áudios e vídeos. Várias estruturas de indexação também foram estudadas para podermos agilizar essa consulta, a que melhor se destacou até o momento foram os grafos de proximidade. O objetivo principal desse trabalho é formar um conjunto de dados que seja capaz de mostrar a diversidade dos conjuntos utilizados na área de recuperação de dados por similaridade, verificando suas características e as comparando com as características extraídas dos outros conjuntos de dados. Além disso, demonstrar o método que foi utilizado para chegar nesse conjunto final. Com esse conjunto montado e analisado procura-se mostrar se é possível melhorar ou talvez enriquecer a escolha dos conjuntos de dados utilizados nos artigos de recuperação por similaridade, observando se*

existe alguma tendência nessa escolha como por exemplo os conjuntos mais populares. Para isso primeiro será feito um levantamento dos conjuntos de dados complexos utilizados por pesquisadores da área, após o levantamento analisaremos as propriedades dos próprios conjuntos de dados. No melhor do nosso conhecimento até hoje na literatura essa análise e levantamento de conjuntos de dados levando em consideração as propriedades dos próprios conjuntos ainda não foi feita. Como resultado final espera-se mostrar como os conjuntos de dados se comportam, disponibilizando o conjunto já extraído e analisado juntamente com o método utilizado para pesquisadores da área analisarem seus próprios conjuntos de dados.

1. Introdução

Dados complexos (imagens, vídeos, áudios etc.) são gerados a todo instante no mundo atualmente, a geração desses dados está presente no dia-a-dia, todas as novas experiências e memórias são registradas. Com esse aumento significativo na quantidade de dados as formas de recuperar ou procurar por esses dados dentro de todo o seu conjunto ficou bem mais complicado. Encontrar exatamente o que você procura em uma imensidão de dados não é uma tarefa simples. Para recuperar essas imagens, áudios ou até mesmo vídeos utiliza-se operações de recuperação por similaridade, o objetivo dessas operações é medir a similaridade entre dois elementos.

Para indexar dados complexos e tornar as buscas por similaridade mais ágeis, existem as estruturas de indexação que podem ser divididas em quatro grupos. Métodos baseados em árvore [23, 18, 24], métodos baseados em permutação [1, 17, 6], métodos baseados em *hashing* [11, 25, 27] e métodos baseados em grafos [12, 19, 16, 22]. Trabalhos recentes sugeriram que métodos baseados em grafo se provaram melhores quanto a aproximação de similaridade entre objetos [17, 11, 8].

Na literatura existem diversos conjuntos de dados na área de recuperação por similaridade, e esses conjuntos são utilizados para as mais variadas aplicações. Aplicações que vão de encontrar um rosto semelhante ou igual ao que você está procurando na galeria de fotos, a até identificar câncer no exame médico de pacientes sem a necessidade que um especialista no assunto olhe os exames. Contudo, no melhor do nosso conhecimento, um levantamento e análise em relação a esses conjuntos, que foram utilizados nos artigos da área de recuperação de dados por conteúdo, tendo o enfoque somente nas características extraídas dos próprios conjuntos de dados ainda não foi feito.

O objetivo desse trabalho é analisar os conjuntos de dados utilizados por pesquisadores da área de recuperação por similaridade. Essa análise será feita baseada somente nas próprias características extraídas dos conjuntos, comparando conjunto a conjunto em grupos e individualmente. Para isso pretende-se fazer um levantamento desses conjuntos de dados, obtendo os conjuntos e juntamente com eles o motivo por trás dos projetos que os utilizaram. Após o levantamento será feita a análise, que seria uma bateria de testes, tendo como resultado o quanto esses conjuntos se aproximam/diferenciam, suas particularidades e seus aspectos principais.

Dito isso a análise como um todo seria interessante pois contribuiria para entender e visualizar a diversidade dos conjuntos de dados que são utilizados nos principais trabalhos de pesquisa por similaridade. Com essa análise seria possível dizer se é possível

enriquecer a escolha dos conjuntos de dados que estão sendo feitas até então nos trabalhos relacionados a similaridade. Outra contribuição seria uma base de dados, onde está claro as características dos conjuntos, facilitando para pesquisadores da área escolher qual melhor se encaixa no problema que estão propondo resolver. Caso os pesquisadores não estejam interessados nos conjuntos disponibilizados, também terão o método que que foi utilizado para gerar o conjunto pra poderem aplicar em seus próprios conjuntos de dados.

A organização deste documento fora realizado da seguinte maneira: Na seção 2 apresenta os fundamentos necessários para o entendimento deste trabalho. Na seção 3 será mostrado o que já existe na literatura para fundamentá-lo. Na seção 4 haverá a proposta de solução junto da descrição do problema e dos objetivos. Na seção 5 haverá os resultados preliminares. Na seção 6 será apresentado como esses objetivos serão atingidos.

2. Fundamentação Teórico-Metodológica e Estado da Arte

Essa seção nos apresenta de uma maneira brave o que são pesquisas por similaridade, pois são em artigos relacionados a elas que estamos interessados. Também apresentaremos uma descrição de o que seriam conjuntos de dados, pois o trabalho como um todo gira em torno de analisar esses conjuntos.

2.1. Pesquisas por Similaridade

Buscas por similaridade são pesquisas que buscam a similaridade entre objetos para retornar resultados. Esse tipo de pesquisa é comum quando procura-se identificar a similaridade entre dados complexos. Dados complexos são dados ricos em conteúdo, como por exemplo imagens, vídeos e áudios. A representação desses dados normalmente se da por meio de vetores de características. Vetores de características são vetores numéricos que são extraídos desses dados complexos (imagens, vídeos, áudios etc..), sendo eles comparados quando ocorre as buscas por similaridade.

Dados complexos não podem ser comparados por operadores comuns como $<$, $>$, $\leq$, $\geq$. Não há a possibilidade de fazer imagem $<$ imagem e obter um resultado satisfatório. A Figura 1[1] apresenta um exemplo de consulta por similaridade, essas consultas podem ter um ou mais elementos de referência. Disponibilizando um elemento de consulta como entrada, recebe-se como retorno elementos de consulta iguais ou semelhante ao elemento disponibilizado

Consultas por similaridade tipicamente dependem de funções de distância que mensuram a dissimilaridade entre elementos de um determinado domínio de dados. O conjunto de vetores de características juntamente com função de distância forma o chamado espaço de similaridade. Esse espaço diz se o objeto em questão está dentro do mesmo espaço de similaridade do objeto buscado. Para aprofundar-se no assunto refira-se as referências [5] e [26].

2.2. Conjuntos de dados

Conjuntos de dados são utilizados em diversas aplicações. Os tipos de dados desses conjuntos podem variar entre imagens, textos, planilhas, áudios, vídeos etc. Esses conjuntos podem ser formados pelos dados propriamente ditos (como exemplo a imagem em si), ou

[1] https://miro.medium.com/max/814/1*26NxutQtQLIGcXNVO4HP3w.png

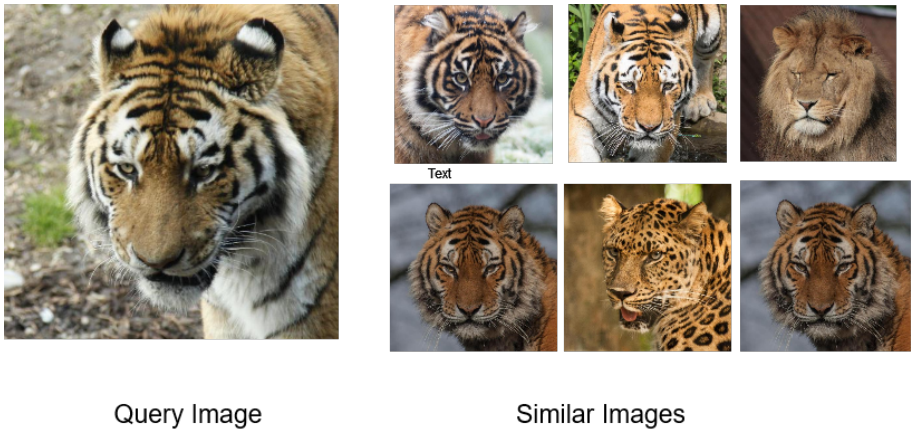

Figura 1. Exemplo prático retornando elementos semelhantes

por meio de vetores de características. Um exemplo de conjunto formado pelo dado propriamente dito é o *SISAP Spanish*[2], que é um conjunto de dados em que seus elementos são palavras de um dicionário de espanhol.

Outro conjuntos podem ser formados por vetores de características, que foram obtidos através de algum método de extração. Esse método vária dependendo do tipo de dados e do objetivo por trás do trabalho que vai utilizar o conjunto. Um exemplo de extrator seria o SIFT [15]. Esse extrator pega os pontos principais das imagens, formando vetores de características numéricos de 128 dimensões para dos pontos.

3. Trabalhos correlatos

Quando se trata de conjuntos de dados, na literatura tem-se artigos relacionados a sua criação [4] [15]. Criar novos conjuntos de dados pode se dar por juntar os dados propriamente ditos(como por exemplo coletar imagens) ou extrair características desses dados(imagens, vídeos, áudios) com um método de extração. Tem-se também artigos que buscam avaliar estruturas de indexação e fazem isso utilizando conjuntos de dados [2]. Por fim tem-se artigos que em vez de comparar elemento a elemento de um conjunto de dados, fazem a comparação conjunto a conjunto, para verificar suas semelhanças [20] [21].

Existem diferentes motivos para criar conjuntos de dados. Um deles seria quando ainda não há um conjunto de dados adequado para o trabalho em questão. Por exemplo tem-se o conjuntos de dados CoPhIR [4] que foi desenvolvido para fazer testes de escalabilidade para pesquisa de similaridade. Outro exemplo de criação de conjunto de dados tem-se o ANN SIFT1B [13] que criou um conjunto de dados a partir de aproximadamente 1 milhão de imagens utilizando o método de extração SIFT [15]. Dependendo do método de extração utilizado, é gerado um conjunto de dados com diferentes características.

Sobre trabalhos da área que avaliam estruturas e métodos de indexação pode-se citar o famoso ANN-Benchmarks [2] que foi criado para medir o quanto a performance dos algoritmos de vizinho mais próximos se aproximam na memória. Nesse trabalho foi-se utilizado 6 diferentes conjuntos de dados (SIFT, Glove, NYTimes, Rand-Angular, SIFT-Hamming e NYTimes-Hamming). Todos esse conjuntos contribuiram no artigo

[2] `http://sisap.org`

respondendo perguntas como performance, o quão robusto era o algoritmo, a qualidade da resposta e diversas combinações possíveis dentro do conjunto com suas métricas de distância.

Pesquisas por similaridade, na maioria das vezes estão relacionadas a pesquisas elemento a elemento. Porém o foco desse trabalho não são pesquisas individuais, mas sim a comparação entre conjunto de dados inteiros. Busca-se entender as características dos conjuntos e como eles se comportam. Como exemplo da comparação entre conjuntos inteiros na literatura tem-se [20], ele mostra um método para medir a similaridade entre conjunto de dados, para isso eles apresentam e avaliam uma nova métrica de similaridade, com essa métrica eles conseguem agrupar conjuntos de dados semelhantes para realizar a mineração de dados. Outro exemplo é [21], onde ele apresenta as similaridades entre conjuntos de dados utilizando um método de dois passos. Primeiro construir um modelo condensado do conjunto de dados, isso envolve encontrar os componentes do modelo e os relacionamentos entre esses componentes. Nesse caso os modelos são os subespaços. E o segundo passo seria identificar semelhanças entre os modelos condensados.

No melhor do nosso conhecimento este trabalho se diferencia dos trabalhos citados pois ele caracteriza os conjuntos de dados de uma maneira diferente. São extraídas características dos conjuntos de dados completos e transformados em um único vetor de características. Essas características são métricas que foram estudadas e geradas por [22] mostrando como melhor descrever um conjunto de dados pelas próprias características dos mesmos. Até hoje ainda não se foi feita uma análise tão profunda na comparação dos conjuntos de dados tomando em conta somente suas características. Outra novidade do do trabalho é que pretende-se mostrar se existem features interdependentes em um ambiente rico em características. E por fim esse projeto trabalha com uma quantidade e variedade de conjuntos de dados muito maior do que tem-se na literatura até então para esse mesmo propósito.

4. Proposta

Como foi dito acima a proposta desse trabalho é analisar os conjuntos de dados complexos usados por pesquisadores da área de recuperação de dados por similaridade. Com essa análise será possível observar o quanto os conjuntos de dados complexos são diversos entre si, e isso com base somente em suas próprias características. A variabilidade dos conjuntos juntamente com o número de conjuntos para determinado artigo são escolhas que é proposto aprimorar com uma análise mais a fundo sobre esses conjuntos.

Em [22], foi feita uma análise mostrando quais as propriedades mais desafiadoras para a recuperação de dados por similaridade. Nele foram utilizadas varias métricas citadas na literatura que exploravam propriedades para medir a complexidade dos conjuntos de dados em relação à recuperação de similaridade, tem-se como exemplo dessas propriedades a dimensionalidade intrínseca [14, 10, 3], o fenômeno da concentração de distâncias [7], e o contraste relativo [9], todas elas foram aplicadas ao problema.

Foi utilizando essas propriedades que foram extraídas as características dos conjuntos de dados e de subconjuntos dos mesmos. Foram feitas análises para ver o quanto os conjuntos se pareciam entre si, tudo isso a partir de suas características. Com essas propriedades extraídas foi possível caracterizar os conjuntos de dados com o método [22]. Outra análise que seria pra ver se todas as features utilizadas estão acrescentando algo as

características dos conjuntos, quais features mais se encaixam nos conjuntos de dados e quais delas são interdependentes.

4.1. Definição do problema

No melhor do nosso conhecimento, na literatura ainda não há um levantamento e análise de conjuntos de dados, levando em conta somente as características dos próprios conjuntos. Os trabalhos da área têm seus conjuntos escolhidos por seus próprios autores, e essa escolha muitas vezes é guiada pela experiência. Sem essa análise pode ser que ocorra alguma tendência na escolha desses conjuntos, como por exemplo por popularidade.

Outro problema é como analisar conjuntos de dados. Quais características melhor os descrevem e por que essas características tem toda essa influência. Sabe-se até então que com o aumento da dimensionalidade maior a complexidade e o tempo de execução. Porém deve-se analisar quais características são interdependentes e quais características melhor descrevem quais tipos de conjuntos. Com os conjuntos de dados juntos, ainda não se sabe como um ambiente rico em características se comporta.

4.2. Objetivos

O objetivo geral é criação e análise de um conjunto de dados que seria a junção dos conjuntos utilizados até então nos trabalhos relacionados a pesquisas por similaridade. E além disso disponibilizar o método para criação desse conjunto.

Objetivos específicos:

- Descobrir alguma tendência na escolha dos conjuntos de dados.
- Enriquecer a escolha de conjuntos de dados dos pesquisadores da área.
- Compreender como as características dos conjuntos dependem umas das outras.
- Compreender como as características impactam nos conjuntos.

5. Resultados preliminares

O começo desse trabalho já foi desenvolvido, um levantamento inicial foi feito para o início das análises. Com esse levantamento tem-se uma noção dos conjuntos básicos utilizados por pesquisadores da área de recuperação por similaridade. Aqui mostra-se alguns resultados que foram obtidos analisando 115 artigos da área. Esses artigos foram publicados pelas conferências SISAP, VLDB e SIGMOD.

O critério para escolha dos artigos se deu se ele falavam ou não de pesquisas por similaridade, caso o artigo fosse em relação a isso ou a algum derivado próximo o mesmo era analisado. Com esse pré-levantamento inicial conseguiu-se alguns resultados interessantes. Como pode ser visto na Figura 2 a maior parte dos artigos analisados foram da SISAP, o que já era esperado pois ela ela é um conferência focada em recuperação de dados por similaridade.

Agora na Figura 3 é demonstrado o número de artigos por ano e por conferência. Na Figura 4 fez-se um world cloud com o nome dos conjuntos de dados utilizados por esses artigos. Nela consegue-se ver muito bem que a ocorrencia de alguns conjuntos de dados é muito maior sobre os outros. Como exemplo disso é possivel ver alguns conjuntos da SISAP como o *Colors*, *English* e *NASA*. Outras ocorrências populares são o *ALOI*, *CoPhIR* e o famoso *MNIST*.

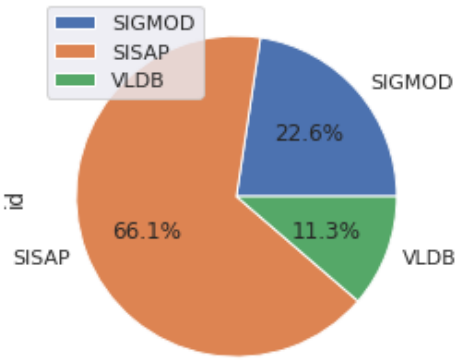

Figura 2. Percentual de artigos por conferência

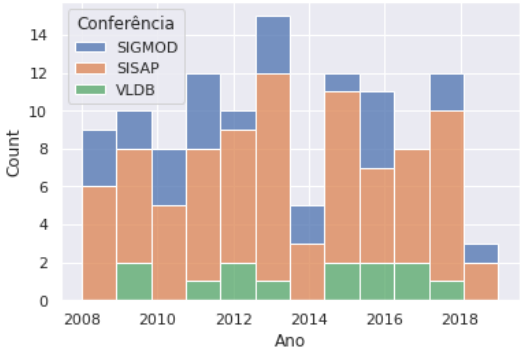

Figura 3. Número de artigos por ano e conferência

Agora levando em conta o conjunto de dados para meta-aprendizado de máquina, construído em conjunto com [22] é visível na Figura 5 como os conjuntos de dados estão distribuídos no espaço definido pelos dois primeiros componentes principais das características extraídas dos conjuntos de dados, mostrando a variabilidade deles no espaço. Como pode-se ver na figura, o conjunto contempla 16 conjuntos de dados, incluindo vários subconjuntos dos mesmos, pois a cardinalidade é um fator determinante para o desempenho de estruturas de indexação. Portando, há conjuntos de várias cardinalidades e dimensionalidades. Os pontos que estão com as mesmas cores são os mesmos conjuntos de dados, porém com alguma variação de cardinalidade e/ou dimensionalidade, mostrando como essa variação impacta na distribuição dos conjuntos de dados no espaço de suas características.

Algumas análises adicionais foram feitas em subconjuntos do conjunto principal, pois como foi dito nos objetivos dependendo dos dados que estão no espaço no momento a análise pode ser diferente. Esse exemplo de análise foi feita a comparação entre os conjuntos SIFT. A semelhança entre eles é que ambos utilizam extratores *SIFT* semelhantes. O primeiro utiliza o extrator *SIFT* padrão, que considera os pontos principais da imagem e gera um vetor de 128 dimensões para cada um desses pontos. Já o segundo, chamado *Dense SIFT*, em vez de identificar os pontos principais, considera todos os pontos da imagem inteira como referência, gerando também vetores de 128 dimensões para cada ponto. A cardinalidade e dimensionalidade utilizadas nessa análise de subconjuntos foram as

Figura 4. WordCloud com o nome dos conjuntos de dados

mesmas. Esse comparativo foi feito para analisar a diferença entre o extrator *SIFT* e o extrator *Dense SIFT*, quando aplicados ao mesmo conjunto de imagens e sob imagens diferentes. Portando foram utilizados o mesmo conjunto de imagens para a extração de ambos.

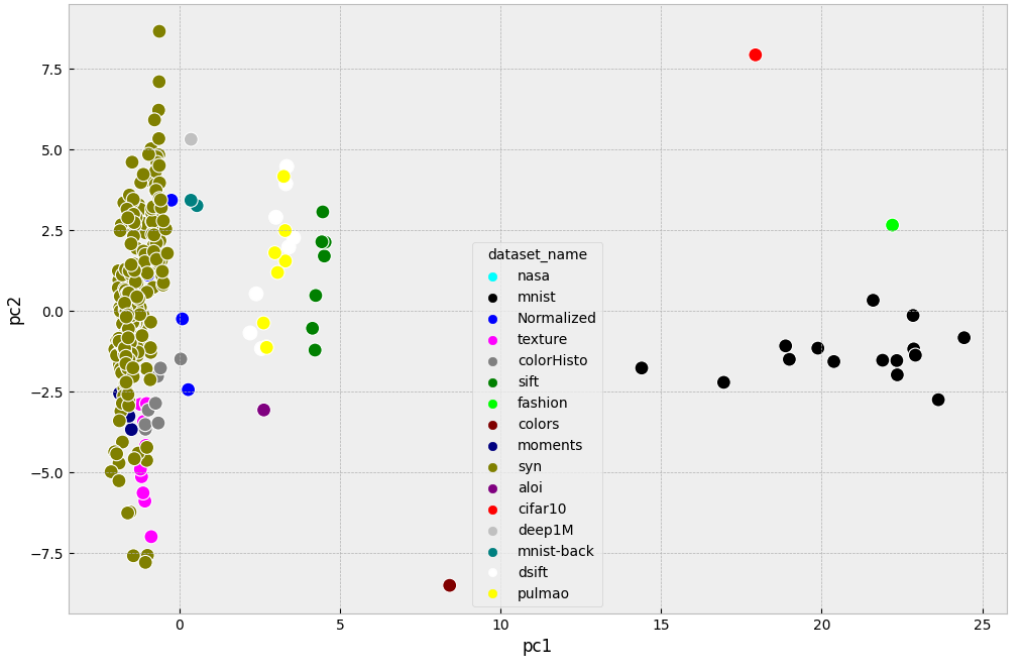

Figura 5. PCA meta conjunto de dados

O objetivo era mostrar que mesmo sendo semelhantes, os extratores não nos mostravam os mesmos resultados. A pergunta a ser respondida aqui seria se vale a pena a utilização de ambos os conjuntos de dados em um experimento ou se seria apenas ”mais do mesmo”. Nesta análise, foram utilizadas duas bases de dados, uma base retirado do site Flickr e a outra de imagens médicas de pulmão. Foram extraídas features *SIFT* e *Dense SIFT* de ambas as bases de imagens, com o propósito de verificar se há diferenças significativas dos features produzidos por cada extrator. Agora na Figura 6 é mostrada que a real diferença entre os *SIFTs* está em seu extrator e não na escolha de imagens.

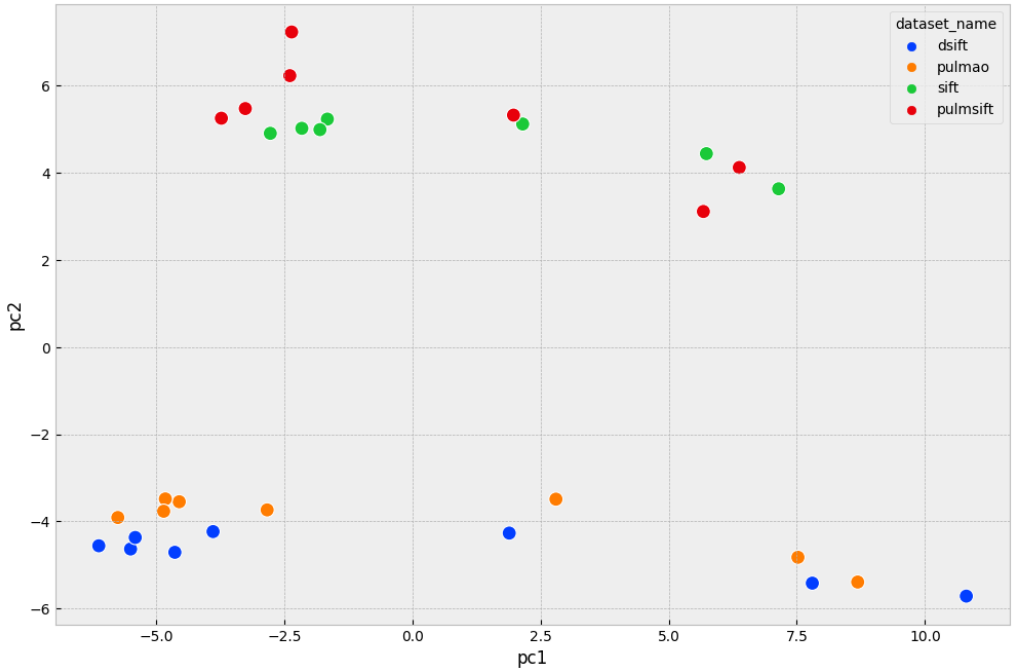

Figura 6. PCA SIFT x DSIFT

6. Procedimentos metodológicos/Métodos e técnicas

Primeiramente deve-se melhorar os pré-resultados, aumentando a integridade da tabela feita para apurar alguns dados que possam ser interessantes para a análise do levantamento inicial. Com esse primeiro levantamento completamente apurado deve-se saber a maioria dos conjuntos de dados utilizados por pesquisadores da área e para que propósito eles estavam utilizando esses conjuntos. Com isso precisa-se começar a busca por esses conjuntos de dados, pois para realizar as análises os conjuntos de dados devem estar em mãos.

Com a busca e obtenção dos conjuntos de dados, deve-se extrair rapidamente o que for necessário e montar os subconjuntos desses mesmos conjuntos (variando dimensionalidade e cardinalidade quando possível). A extração vai ser feita de acordo com o método criado por [22], onde ele demonstra por meio de métricas e meta-aprendizagem a melhor maneira de representar um conjunto de dados. Após a extração das características a primeira tarefa a ser feita para as análises é um estudo para entender a importância dos dessas características e como elas descrevem cada um dos dataset. Esse estudo é importante pois um dos problemas que temos em relação a isso é que algumas características geradas pelo método de [22] vem vazios dependendo do conjuntos. Isso deve ser tratado e analisado. Deve-se entender isso para descobrir qual a melhor maneira de tratar esse acontecimento.

Feita a análise de porque algumas características vem vazias e como tratá-las, outra análise para as características seriam quais delas são mais importantes para determinados conjuntos de dados, vendo o impacto tem na comparação entre conjuntos de dados . Juntamente com o essas características importantes, vamos buscar ver a correlação entre essas características como por exemplo tempo de resposta e qualidade da resposta, porém em um conjunto mais rico em características.

Com isso deve-se testar os conjuntos em grupos separados, pra ver qual comportamento eles têm quando não estão no conjunto como um todo. Um exemplo disso já foi mostrado no teste dos conjuntos *SIFTs*, onde dois conjuntos de dados do conjunto completo eram comparados para mostrar resultados diferentes. Isso deve ser pensado e testado para conseguir mostrar as diferenças entre os conjuntos que quando estavam no conjunto completo pareciam não ter diferença, porém quando separados e testados sob uma outra perspectiva pode ser que sejam bem diferentes. Isso tudo leva em consideração as características dos próprios conjuntos e dos possíveis objetivos que os trabalhos que vão os usar tenham.

7. Cronograma de Execução

As atividades apresentadas na seção 6 serão listadas abaixo e, em seguida, há um cronograma que mostra a ordem de realização destas atividades na Tabela 1.

Atividades:

1. Revisão bibliográfica do estado da arte apurando a tabela de pré-resultados;
2. Obter os conjuntos de dados encontrados e fazer o download dos mesmos;
3. Estudar a proposta feita por [22] para entender melhor os features;
4. Analisar quais as características tem mais impactos sob os resultados;
5. Analisar como as características dependem uma da outra;
6. Construção do conjunto de dados formado por todos os conjuntos fornecidos pelos trabalhos da área de recuperação por similaridade;
7. Analisar como as características desses conjuntos se aproximam/diferenciam;
8. Montar o passo a passo do nosso método para os pesquisadores utilizarem;
9. Escrita do Trabalho de Conclusão de Curso;

Tabela 1. Cronograma de Execução

	set	out	nov	dez	jan	fev	mar	abr	mai	jun
Atividade 1	X	X								
Atividade 2		X	X							
Atividade 3			X	X						
Atividade 4			X	X	X	X				
Atividade 5				X	X	X	X			
Atividade 6					X	X	X	X	X	
Atividade 7						X	X	X	X	X
Atividade 8							X	X	X	X
Atividade 9				X	X	X	X	X	X	X

8. Contribuições e/ou Resultados esperados

Esse trabalho tem como expectativa:

- Mostrar o quão diversos são os conjuntos de dados utilizados na literatura.
- Enriquecer a escolha de conjuntos de dados para pesquisadores da área.
- Identificar se há alguma tendência na escolha dos conjuntos pelos artigos relacionados a similaridade.

- Fornecer um conjunto de dados analisado onde pesquisadores que estiverem interessados nesses conjuntos possam escolher os que melhor se encaixem em seus projetos.
- Dar um método para pesquisadores que não estejam interessados no nosso conjuntos poderem fazer os mesmos passos que os nossos para seus próprios conjuntos de dados.
- Identificar como as features se comportam, quais características elas seguem em um ambiente rico em features.

9. Espaço para assinaturas

Londrina, *13 de setembro de 2021*.

_____ _____

Aluno Orientador

Referências

[1] Giuseppe Amato, Claudio Gennaro, and Pasquale Savino. Mi-file: using inverted files for scalable approximate similarity search. *Multimedia tools and applications*, 71(3):1333–1362, 2014.

[2] Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. In *International Conference on Similarity Search and Applications*, pages 34–49. Springer, 2017.

[3] Martin Aumüller and Matteo Ceccarello. Benchmarking nearest neighbor search: Influence of local intrinsic dimensionality and result diversity in real-world datasets. In *EDML@ SDM*, pages 14–23, 2019.

[4] Paolo Bolettieri, Andrea Esuli, Fabrizio Falchi, Claudio Lucchese, Raffaele Perego, Tommaso Piccioli, and Fausto Rabitti. Cophir: a test collection for content-based image retrieval. *arXiv preprint arXiv:0905.4627*, 2009.

[5] Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. Searching in metric spaces. *ACM computing surveys (CSUR)*, 33(3):273–321, 2001.

[6] Andrea Esuli. Use of permutation prefixes for efficient and scalable approximate similarity search. *Information Processing & Management*, 48(5):889–902, 2012.

[7] Damien François, Vincent Wertz, and Michel Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886, 2007.

[8] Kiana Hajebi, Yasin Abbasi-Yadkori, Hossein Shahbazi, and Hong Zhang. Fast approximate nearest-neighbor search with k-nearest neighbor graph. In *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[9] Junfeng He, Sanjiv Kumar, and Shih-Fu Chang. On the difficulty of nearest neighbor search. *arXiv preprint arXiv:1206.6411*, 2012.

[10] Michael E Houle. Local intrinsic dimensionality i: an extreme-value-theoretic foundation for similarity applications. In *International Conference on Similarity Search and Applications*, pages 64–79. Springer, 2017.

[11] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[12] Jerzy W Jaromczyk and Godfried T Toussaint. Relative neighborhood graphs and their relatives. *Proceedings of the IEEE*, 80(9):1502–1517, 1992.

[13] Hervé Jégou, Romain Tavenard, Matthijs Douze, and Laurent Amsaleg. Searching in one billion vectors: re-rank with source coding. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 861–864. IEEE, 2011.

[14] Flip Korn, B-U Pagel, and Christos Faloutsos. On the"dimensionality curse"and the"self-similarity blessing". *IEEE Transactions on Knowledge and Data Engineering*, 13(1):96–111, 2001.

[15] David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

[16] Yury Malkov, Alexander Ponomarenko, Andrey Logvinov, and Vladimir Krylov. Approximate nearest neighbor algorithm based on navigable small world graphs. *Information Systems*, 45:61–68, 2014.

[17] Bilegsaikhan Naidan, Leonid Boytsov, and Eric Nyberg. Permutation search methods are efficient, yet faster search is possible. *arXiv preprint arXiv:1506.03163*, 2015.

[18] Gonzalo Navarro and Nora Reyes. Dynamic spatial approximation trees for massive data. In *2009 Second International Workshop on Similarity Search and Applications*, pages 81–88. IEEE, 2009.

[19] Rodrigo Paredes and Edgar Chávez. Using the k-nearest neighbor graph for proximity searching in metric spaces. In *International Symposium on String Processing and Information Retrieval*, pages 127–138. Springer, 2005.

[20] Srinivasan Parthasarathy and Mitsunori Ogihara. Exploiting dataset similarity for distributed mining. In *International Parallel and Distributed Processing Symposium*, pages 399–406. Springer, 2000.

[21] Karlton Sequeira and Mohammed J Zaki. Exploring similarities across high-dimensional datasets. In *Research and Trends in Data Mining Technologies and Applications*, pages 53–84. IGI Global, 2007.

[22] Larissa C Shimomura, Rafael Seidi Oyamada, Marcos R Vieira, and Daniel S Kaster. A survey on graph-based methods for similarity searches in metric spaces. *Information Systems*, page 101507, 2020.

[23] Caetano Traina, Agma Traina, Bernhard Seeger, and Christos Faloutsos. Slim-trees: High performance metric trees minimizing overlap between nodes. In *International Conference on Extending Database Technology*, pages 51–65. Springer, 2000.

[24] Marcos R Vieira, Caetano Traina Jr, Fabio JT Chino, and Agma JM Traina. Dbm-tree: A dynamic metric access method sensitive to local density data. *Journal of Information and Data Management*, 1(1):111–111, 2010.

[25] J Wang, HT Shen, J Song, and J Ji. Hashing for similarity search: A survey, corr abs/1408.2927. *arXiv preprint arXiv:1408.2927*, 2014.

[26] Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. *Similarity search: the metric space approach*, volume 32. Springer Science & Business Media, 2006.

[27] Yan-Ming Zhang, Kaizhu Huang, Guanggang Geng, and Cheng-Lin Liu. Fast knn graph construction with locality sensitive hashing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 660–674. Springer, 2013.