Análise da Evolução Temporal de Dados Complexos

Isis Caroline O. V. de Sousa, Renato Bueno

Programa de Pós Graduação em Ciência da Computação

Departamento de Computação – Universidade Federal de São Carlos (UFSCar)

São Carlos – SP – Brasil

{isis.sousa, renato} @dc.ufscar.br

Nível: Mestrado

Ingresso no programa: Primeiro semestre de 2012

Exame de qualificação: junho de 2013

Época esperada de conclusão: março de 2014

Abstract. *In complex data, it is common to make similarity queries, using the features extracted of these data. These data are in general represented in metric spaces, where only the elements and their features are known. Considering the necessity of associate time to metric data, the objective is analyze the temporal evolution of the metric-temporal data, proposed on a previous work, through the embedding of these data to multidimensional spaces. In the multidimensional space, we can analyze the trajectories, estimate the data's status in different moments of time and perform similarity search to this estimate in the multidimensional space. Initially, we intend to study algorithms of embedding that can preserve the distances between the elements as in the original space. Then, we will propose and evaluate different kinds of similarity search to perform estimates in the embedding space evaluating the outcomes in the moment of the search, starting with Range Query and Reverse k -NN. To finish, we intend to evaluate the use of multiple reference elements, when they are available, to calculate the estimates. With the results of the embedding and the queries on the estimates, we intend to validate the method proposed providing a support to real applications.*

Keywords: Complex Data. Metric Space. Temporal Evolution. Embedding. Similarity Search.

1. Problema de pesquisa e caracterização da contribuição

Além dos tipos convencionais de dados (números, datas e pequenas cadeias de caracteres), é cada vez mais comum a necessidade de suportar dados complexos, como imagens, áudio, vídeos, etc. Para realizar uma consulta aos dados complexos, um conjunto de características pode ser extraído dos mesmos e as comparações entre eles são baseados na relação de (dis)similaridade entre seus vetores de características [9]. Dessa forma, o problema essencial é encontrar, no conjunto de elementos, aqueles que são mais similares ao elemento de consulta, utilizando uma função de distância [1]. Realiza-se então consultas por similaridade, utilizando os sistemas baseados em Recuperação por Conteúdo (*Contend-based Retrieval* – CBR). Em se tratando de imagens (tipo de dado complexo a ser utilizado neste trabalho), utiliza-se a Recuperação de Imagem por Conteúdo (*Content-based Image Retrieval* – CBIR). Dados complexos são em geral representados em espaços métricos, domínio de dados onde as únicas informações disponíveis são os elementos e as distâncias entre eles.

A principal motivação para o estudo proposto é a necessidade de acrescentar a informação temporal aos dados métricos, pois em muitas aplicações, como acompanhamento de pacientes através de imagens de exames médicos, a associação do tempo é fundamental. Em [3] foi proposto o espaço métrico-temporal, que consiste no acréscimo da informação do tempo no espaço métrico, através de uma componente temporal. Outra proposta preliminar, idealizada por [2], refere-se ao mapeamento do espaço métrico-temporal para um espaço multidimensional, possibilitando, com os dados representados nesse espaço, analisar seu comportamento evolutivo no decorrer do tempo.

O objetivo deste trabalho é aperfeiçoar e estender a proposta preliminar de [2] para análise da evolução temporal de dados complexos. Pretende-se inicialmente escolher algoritmos de mapeamento dos dados para o espaço multidimensional que mantenham a distribuição dos dados no espaço original, possibilitando uma melhor avaliação da qualidade das estimativas. Serão propostas novas maneiras de se estimar o estado de um elemento em momentos de tempo diferentes daqueles em que estão disponíveis na base de dados. Por fim, será avaliada a utilização de múltiplos elementos de referência, em diferentes momentos, estendendo assim a proposta original.

O restante do artigo está estruturado da seguinte maneira: Na seção 2 é apresentada a fundamentação teórica do problema de pesquisa. Na seção 3 é apresentada a proposta preliminar de [2]. Na seção 4 é apresentado o desenvolvimento necessário para conclusão e na seção 5 as considerações finais e os resultados esperados.

2. Fundamentação Teórica

Dados complexos, diferentemente dos convencionais, não possuem uma Relação de Ordem Total, impossibilitando a utilização dos operadores relacionais ($<$, $\leq$, $>$, $\geq$). O procedimento comumente adotado é representar esses dados em espaços métricos e recuperá-los por similaridade.

Nas consultas por similaridade, o conteúdo do objeto complexo é comumente representado através do seu vetor de características. Os vetores de características são comparados para que se possa obter o grau de similaridade entre os objetos, onde quanto menor for o valor da distância resultante, maior é a similaridade entre os objetos.

Os tipos de consulta por similaridade mais comum são: (a) *Range Query*: a partir de um objeto de consulta e um raio de abrangência, são retornados os objetos que estão dentro desse raio; e (b) *k-Nearest Neighbor (k*-NN*)*: a partir de um objeto de consulta e um número k de elementos que devem ser recuperados, são retornados os k mais próximos. Dentre vários outros métodos e variações, neste trabalho pretende-se utilizar a *Reverse k*-NN [10] para avaliar a qualidade das estimativas retornadas por uma consulta. Em uma consulta *Reverse k*-NN são retornados os elementos que tem o objeto de consulta como um dos *k* vizinhos mais próximos.

2.1 Espaços Métricos

Em um espaço métrico não são consideradas informações geométricas ou dimensionais dos dados, isto é, só estão disponíveis os elementos de dados e as distâncias (dissimilaridades) entre eles. Para calcular as distâncias entre os elementos do espaço, a função de distância (métrica) deve satisfazer três propriedades [4]: Simetria ($d(x,y) = d(y,x)$), Não-negatividade ($0 < d(x,y) < \infty,\ x \neq y, d(x,x) = 0$), e Desigualdade Triangular ($d(x,y) \leq d(x,z) + d(z,y)$).

Para incorporar a informação temporal nas consultas por similaridade, foi desenvolvido o espaço métrico-temporal [3], composto por dois espaços métricos (uma componente métrica e uma temporal), sendo: $\langle S, d_s \rangle$ o espaço métrico para o cálculo da similaridade, sendo que S representa o conjunto de dados e d_s uma métrica para calcular a similaridade entre os elementos; e $\langle T, d_t \rangle$ o espaço métrico para as medidas de tempo, em que T representa as medidas de tempo e d_t a métrica para o cálculo da similaridade entre os valores de tempo, que podem ser instantes ou períodos temporais.

3. Proposta Preliminar

As consultas realizadas em espaços métricos ou mesmo métrico-temporais não permitem analisar a evolução temporal dos dados, pois mesmo com as informações temporais presentes, são disponíveis somente os elementos e as distâncias entre eles. Essa análise é necessária em muitos domínios de aplicação, para analisar o comportamento evolutivo dos dados complexos no decorrer do tempo, como na medicina, meteorologia, agricultura, entre outras. Um exemplo aplicado à medicina é o acompanhamento do diagnóstico de um paciente através de imagens de exames médicos.

Para estudar essa necessidade, pretende-se estender, nesse trabalho, a proposta apresentada por [2], onde foi proposto o mapeamento do espaço métrico-temporal para um espaço multidimensional, para possibilitar a análise da evolução temporal dos dados complexos. Baseando-se nas informações existentes de um determinado objeto no banco de dados (instâncias desse objeto em tempos diferentes), pode-se estimar e analisar o estado desse mesmo objeto em um outro instante no tempo.

Considere como exemplo o acompanhamento do diagnóstico de um paciente através de imagens de exames médicos, como ilustrado na Figura 1. No espaço mapeado, os pontos representam as imagens mapeadas de exames de pacientes. Existem duas imagens indexadas referentes ao paciente P_A, uma antes de iniciar o tratamento $(t = 0)$ e outra com 12 meses de tratamento $(t = 12)$. Deseja-se estimar o estado desse paciente quando ele estiver com 15 meses de tratamento. Por meio das posições das duas imagens de P_A existentes (em $t = 0$ e $t = 12$), estima-se qual será sua posição em

$t = 15$, utilizando interpolação/extrapolação. Porém, não é possível a construção/reconstrução de uma imagem (no espaço original) a partir dessa estimativa no espaço multidimensional. Realiza-se então uma consulta por similaridade (k-NN) no espaço multidimensional utilizando essa posição estimada como centro de consulta. Os objetos retornados são aqueles presentes na base que são os mais próximos da estimativa de P_A em $t = 15$.

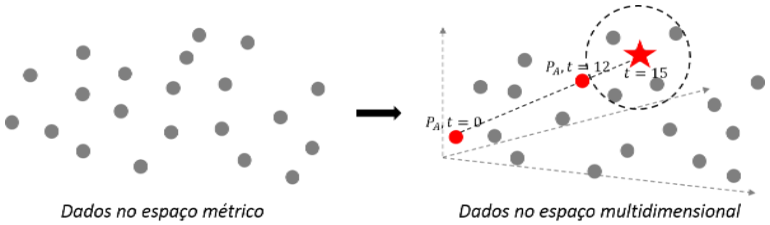

Figura 1: Exemplo de estimativa e consulta no espaço mapeado.

3.1 Experimentos

Em [2] foram realizados experimentos utilizando o algoritmo *FastMap* [6] para o mapeamento dos dados métricos. Foi utilizado o conjunto de imagens ALOI (*Amsterdam Library of Object Images*) [7], onde cada objeto (de um conjunto de 1000 objetos) foi fotografada em 72 ângulos de visão (com rotação de 5 graus entre uma e outra). Admite-se que diferentes posições dos objetos referem-se a diferentes tempos, como mostrado na Figura 2.

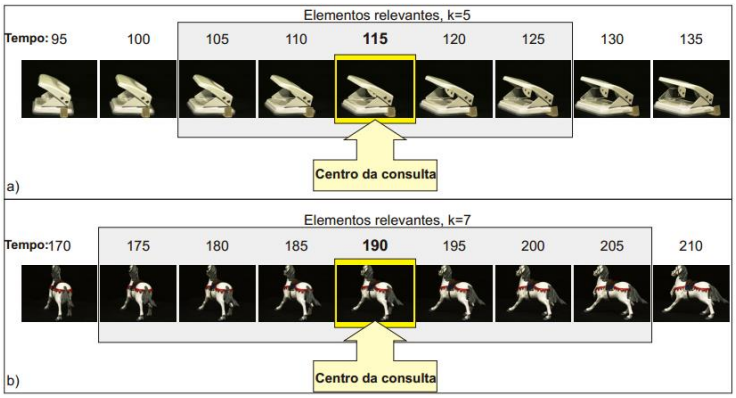

Figura 2: Exemplo do conjunto de imagens utilizado. [2].

Para cada objeto foram utilizadas duas imagens em tempos diferentes, a partir das quais foi estimado o estado do objeto de consulta em outra posição temporal. Foram extraídas características de cor (Histogramas) e formas (Zernike). Por exemplo, a partir de duas instâncias de um objeto x nos tempos 5 e 15, realizou-se uma estimativa da posição do objeto no tempo 20 e realizou-se uma consulta k-NN nessa posição. Para verificar a precisão da consulta, foi realizada, no espaço métrico-temporal original, a mesma consulta k-NN utilizando o objeto x no tempo 20 e os resultados das consultas foram comparados. Essa comparação foi utilizada para avaliar a qualidade das estimativas, realizadas em tempo passado, intermediário e futuro.

Para avaliar a qualidade do mapeamento, como mostrado no gráfico apresentado na Figura 3, a curva denominada "exato" compara os resultados de consultas k-NN sobre objetos no espaço métrico original com os resultados das mesmas consultas realizadas diretamente no espaço mapeado. As demais curvas indicam os resultados

referentes à avaliação das estimativas. Como pode ser visto, a qualidade dos resultados na análise das estimativas teve comportamento muito parecido com a qualidade do mapeamento, sendo um forte indicativo de que, por terem os mesmo níveis de precisão, a qualidade do mapeamento influenciou na qualidade das consultas às estimativas.

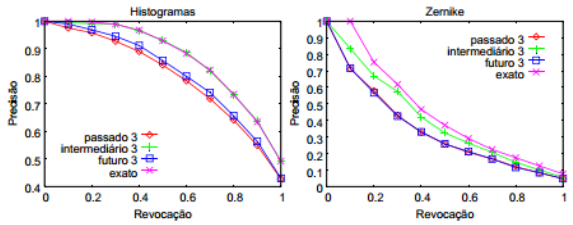

Figura 3: Avaliação da qualidade do mapeamento [2].

4. Desenvolvimento necessário para conclusão

Para dar continuidade ao estudo apresentado em [2], esse trabalho tem como parte da proposta a utilização de outros algoritmos para fazer o mapeamento dos dados do espaço métrico-temporal para o espaço multidimensional, a fim de manter o mais fielmente possível a distribuição dos dados. Considerando a possibilidade de maior precisão, dois algoritmos foram pré-selecionados a partir da literatura para mapear os dados e comparar os resultados, inclusive com o algoritmo inicialmente utilizado (*FastMap*). O algoritmo MDS clássico [11], técnica referente ao *Multidimensional Scaling* (MDS) faz o mapeamento a partir de uma matriz de distâncias entre os dados e a decomposição espectral dessa matriz. O *SparseMap* [8] faz o mapeamento dos dados a partir dos cálculos de distância entre os objetos do conjunto inicial e seus subconjuntos. Para testar os algoritmos, serão utilizadas bases de dados controladas considerando tamanho da base, custo do mapeamento, precisão e distribuição dos dados.

Numa segunda etapa da proposta, serão estudados outros tipos de consulta que possam proporcionar melhorias nos resultados e nas avaliações dos mesmos. As consultas propostas em [2], realizadas apenas com *k*-NN, impossibilitam a avaliação da qualidade dos resultados no momento da consulta. Além disso, os vizinhos mais próximos retornados podem não ser próximos o suficiente da estimativa para representar um bom resultado. Por exemplo, deseja-se estimar o estado de dois pacientes P_A e P_B nos tempos 15 e 8, respectivamente, como pode ser visto na Figura 4. Os objetos retornados para P_A em $t = 15$ estão muito mais distantes da estimativa de P_A em $t = 15$ do que aqueles retornados para P_B em $t = 8$ estão da estimativa de P_B em $t = 8$. Logo, a qualidade dos elementos retornados para P_B tende a ser superior à qualidade dos que foram retornados para P_A, pois provavelmente serão imagens mais próximas do que se estima ser o estado do paciente P_B com 8 meses de tratamento.

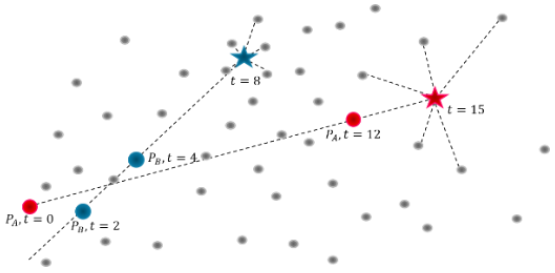

Figura 4: Exemplo de estimativa utilizando *k*-NN.

Serão estudadas e propostas duas possíveis maneiras de avaliar a qualidade das estimativas no momento da consulta, utilizando *Range Query* e *Reverse k*-NN.

Utilizando a *Range Query*, é possível delimitar a distância máxima desejada para os objetos retornados, sendo o raio dessa consulta um indicativo da qualidade dos resultados da estimativa. Isto é, dado um determinado raio de distância, se não houver elementos nesse raio, elementos mais distantes não serão retornados, diminuindo de certa forma a incidência de falsos positivos. Pretende-se estudar maneiras de utilizar a dimensão intrínseca do conjunto para a definição desse raio de abrangência, sendo que a dimensão intrínseca de um conjunto de dados é a dimensão do objeto no espaço representado pelo conjunto, independentemente do espaço onde eles estão imersos [5].

Utilizando a consulta *Reverse k*-NN após a *k*-NN, pode-se conferir se os objetos retornados possuem o objeto de consulta como um dos vizinhos mais próximos. Se os objetos retornados não possuírem o objeto de consulta como um de seus vizinhos mais próximos, pode ser um indicativo de que esse objeto retornado não representa um bom resultado.

Em uma terceira etapa do trabalho, pretende-se estudar maneiras de realizar as estimativas utilizando um número maior de elementos de referência, sendo que na proposta de [2] foram utilizados apenas dois elementos, com estimativas utilizando apenas interpolação/extrapolação. Quando presentes na base de dados, mais objetos podem ser utilizados com parâmetro para fazer a estimativa no tempo desejado. Além disso, pretende-se comparar os resultados das consultas realizadas a partir de dois objetos e a partir de um número maior deles, a fim de analisar se os resultados apresentam melhorias em precisão, considerando também o custo computacional adicional. No exemplo ilustrado na Figura 5, onde deseja-se também estimar o estado de P_A no tempo 15 e de P_B no tempo 8, se houver mais de duas imagens referentes a cada objeto, elas podem também ser utilizadas para fazer as estimativas.

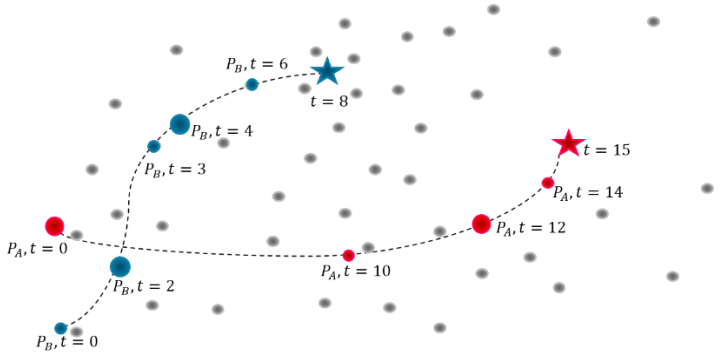

Figura 5: Exemplo de estimativa utilizando mais de dois objetos de referência.

5. Considerações Finais e Resultados Esperados

Considerando que não é possível fazer uma análise da evolução temporal em dados puramente métricos, o mapeamento desses dados para um espaço multidimensional torna-se necessário. Mesmo no espaço métrico-temporal, embora exista a informação do tempo, essa informação não representa necessariamente uma ordem cronológica. Verificada a efetividade da análise das trajetórias dos dados métrico-temporais, pode-se então proporcionar um modelo de aplicação que auxilie muitos domínios de aplicação que utilizam dados complexos.

Logo, espera-se, como resultados desse trabalho:

- Escolher novos métodos de mapeamento dos espaços métricos para espaços multidimensionais, visando manter a distribuição original dos dados, para que seja possível melhor avaliar a qualidade das estimativas;
- Estudar maneiras de avaliar a qualidade das respostas na consulta às estimativas no momento da consulta; e
- Explorar a utilização de mais elementos de referência para gerar as estimativas.

Com os resultados, espera-se possibilitar o uso desse modelo de análise dos dados complexos a aplicações reais, auxiliando nas aplicações necessárias.

6. Referências

[1] ALMEIDA, J. et al. **DAHC-tree: An Effective Index for Approximate Search in High-Dimensional Metric Spaces.** JIDM, v. 1(3), p. 375-390, 2010.

[2] BUENO, R. **Tratamento do tempo e dinamicidade em dados representados em espaços métricos.** Tese (Doutorado em Ciência da Computação). Instituto de Ciências Matemáticas e de Computação, USP. São Carlos, 2009.

[3] BUENO, R. et al. **Time-Aware Similarity Search: A Metric-Temporal Representation for Complex Data.** Proceedings of the 11th International Symposium on Advances in Spatial and Temporal Databases. Aalborg, Denmark: Springer-Verlag: 302-319 p. 2009.

[4] C. TRAINA, J. et al. **Fast Indexing and Visualization of Metric Data Sets using Slim-Trees.** IEEE Trans. on Knowl. and Data Eng., v. 14, n. 2, p. 244-260, 2002. ISSN 1041-4347.

[5] C. TRAINA, J. et al. **Fast feature selection using fractal dimension.** Brazilian Symposium on Databases (SBBD). MEDEIROS , C. M. B. E. B., K. EDITORS. João Pessoa, PB: p. 158-171, 2000.

[6] FALOUTSOS, C.; LIN, K.-I. **FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets.** SIGMOD Rec., v. 24, n. 2, p. 163-174, 1995. ISSN 0163-5808.

[7] GEUSEBROEK, J.-M.; BURGHOUTS, G. J.; SMEULDERS, A. W. M. **The Amsterdam Library of Object Images.** Int. J. Comput. Vision, v. 61, n. 1, p. 103-112, 2005. ISSN 0920-5691.

[8] HRISTESCU, G.; FARACH-COLTON, M. **Cluster-preserving Embedding of Proteins.** Center for Discrete Mathematics; Theoretical Computer Science. 1999

[9] KASTER, D. S. et al. **Nearest Neighbor Queries with Couting Aggregate-based Conditions.** JIDM, v. 2(3), p. 401-416, 2011.

[10] TAO, Y. et al. **Multidimensional reverse kNN search.** The VLDB Journal, v. 16, n. 3, p. 293-316, 2007. ISSN 1066-8888.

[11] YOUNG, G.; HOUSEHOLDER, A. S. **Discussion of a set of points in terms of their mutual distances**. Psychometrika**,** v. 3, n. 1, p. 19-22, 1938.