ImageDW-index: Uma estratégia de indexação voltada ao processamento de imagens em data warehouses

Jefferson William Teixeira[1],
Profa. Dra. Cristina Dutra de Aguiar Ciferri[1]

[1] Pós-Graduação em Ciências de Computação e Matemática Computacional
Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP)
São Carlos, SP, Brasil

{william,cdac}@icmc.usp.br

Nível: Mestrado
Ano de ingresso no programa: 2012
Exame de qualificação: Abril de 2013
Época esperada de conclusão e defesa: Abril de 2014

Abstract. *A data warehousing environment offers support to the decision-making process. It consolidates data from distributed, autonomous and heterogeneous information sources into one of its main components, the data warehouse. Furthermore, it provides efficient processing of analytical queries (i.e. OLAP queries). A conventional data warehouse stores only alphanumeric data. On the other hand, an image data warehouse stores not only alphanumeric data but also intrinsic features of images, thus allowing non-conventional data warehousing environments to perform OLAP similarity queries over images. This requires the development of strategies to provide efficient processing of these complex and costly queries. In our master's research, we focus on this issue. We are developing the ImageDW-index, an index strategy aimed at the efficient processing of analytical queries extended with image similarity predicates. Although there are a number of approaches in the literature that propose indices for data warehouses and indices for image data separately, to the best of our knowledge, there is not an approach that investigate these two issues in the same setting. Therefore, our master's research aims to investigate this gap in the literature.*

Palavras-Chave*: data warehouse de imagens, consultas por similaridade, consultas OLAP, índices*

1. Introdução

Um ambiente de *data warehousing* (DWing) consolida dados de interesse de provedores de informação autônomos, distribuídos e heterogêneos em único banco de dados, o *data warehouse* (DW). Esse ambiente garante eficiência e flexibilidade na recuperação de informações estratégicas voltadas aos processos de gerência e de tomada de decisão [Chaudhuri and Dayal 1997]. Nesse ambiente, consultas analíticas, denominadas OLAP *(on-line analytical processing)*, são executadas diretamente no DW, sem acesso aos provedores originais.

Usualmente a modelagem de um DW segue o esquema estrela, o qual, em um ambiente relacional, consiste de uma tabela de fatos que armazena as medidas numéricas de interesse do DW, bem como referências às várias tabelas de dimensão, as quais contextualizam essas medidas. Por exemplo, em um ambiente de DWing para a área médica, dados relativos à incidência de câncer de mama (medida numérica) podem ser integrados segundo diferentes hospitais e faixas etárias ao longo dos anos (dimensões), oferecendo suporte para a execução de consultas tais como "Qual a incidência de câncer de mama nos últimos três anos em diferentes hospitais, considerando diferentes faixas etárias?".

DWs convencionais armazenam apenas dados alfanuméricos. Entretanto, tem-se estudado a incorporação de dados complexos (imagens) ao ambiente de DWing, de forma a permitir que usuários de sistemas de suporte à decisão (SSD) explorem uma nova gama de consultas analíticas que envolvam comparação de imagens. Por exemplo, certa equipe médica pode estar interessada na seguinte consulta "Qual a quantidade de imagens similares a uma determinada imagem de câncer de mama ocorreram nos últimos três anos no Hospital das Clínicas de Ribeirão Preto?".

Um DW de imagens armazena não somente dados alfanuméricos, mas também dados relacionados a imagens. Por ser uma área de pesquisa recente, ainda não existe um consenso sobre a definição de um DW de imagens com relação ao esquema, aos dados e ao processamento de consultas. Neste artigo, é usada a definição introduzida em [Annibal et al. 2010], a qual considera que um DW de imagens é projetado segundo um esquema estrela diferenciado pois, além de possuir tabelas de dimensão com dados convencionais, esse esquema também possui uma ou mais tabelas de dimensão especificamente voltadas à manipulação de imagens por meio de vetores de características.

Uma questão importante no desenvolvimento desses ambientes não convencionais de DWing refere-se à necessidade de processamento eficiente de consultas analíticas estendidas com predicados de similaridade de imagens. Embora na literatura existam abordagens voltadas à indexação em ambientes de DWing convencionais e à indexação de imagens em bancos de dados complexos, essas propostas apresentam como limitação o fato de não considerarem essas duas áreas de pesquisa conjuntamente.

O projeto de mestrado visa suprir essa limitação existente na literatura, por meio da proposta do ImageDW index, uma estratégia de indexação voltada ao processamento eficiente de consultas analíticas estendidas com predicados de similaridade de imagens. A estratégia em desenvolvimento utiliza conceitos bem difundidos de DW (ex.: índice bitmap de junção) e de armazenamento e recuperação de imagens (ex.: técnica Omni).

Esse artigo está estruturado da seguinte forma. Na seção 2 é descrita a fundamentação teórica, na seção 3 são resumidos os trabalhos correlatos, e na seção 4

é detalhada a proposta do ImageDW-index e descrito o estágio atual de desenvolvimento do trabalho. O artigo é concluído na seção 5, com as considerações finais e próximas atividades a serem desenvolvidas.

2. Fundamentação Teórica

2.1. Índice Bitmap de Junção

Um índice bitmap consiste de vários vetores de bits, cada um construído para um valor do domínio de um atributo. Cada entrada desses vetores faz referência a uma tupla da base de dados e contém o bit "1" se a tupla original possui o valor representado, ou o bit "0", caso contrário [O'Neil and Quass 1997]. Como resultado, operações lógicas bit a bit são realizadas rapidamente pelos processadores. Técnicas de codificação e compressão também são usadas para melhorar o desempenho de índices bitmaps. Ademais, a técnica de *binning* reduz o tamanho do índice criando-se grupos de identificadores pelos quais os valores de um atributo são organizados [Wu et al. 2008].

O índice bitmap é usado em ambientes de DWing convencionais para evitar a necessidade de se realizar operações de junção entre as tabelas de fatos e as tabelas de dimensão no processamento de consultas OLAP. Nesse sentido, para cada atributo de cada tabela de junção, um índice bitmap de junção [O'Neil and Graefe 1995] pode ser construído para indicar o conjunto de tuplas da tabela de fatos que faz junção com os valores daquele atributo. Índices bitmap são adequados a bases de dados do tipo *read-only*, como é o caso de DWs, devido ao alto custo de atualização desse tipo de índice.

2.2. Espaço Métrico e Técnica Omni

Dados complexos podem ser modelados em um espaço métrico, o qual é um par ordenado $(\mathcal{U}, d)$, onde $\mathcal{U}$ é um conjunto de objetos e $d : \mathcal{U} \times \mathcal{U} \rightarrow \mathbb{R}^+$ é uma métrica, isto é, uma função de distância que mede o grau de (dis)similaridade entre os objetos e obedece às propriedades de identidade, simetria, não-negatividade e desigualdade triangular [Ciaccia and Patella 2002]. Nesse espaço, podem ser realizadas consultas por similaridade, dentre as quais tem-se as consultas por abrangência, as quais retornam os elementos contidos em um raio de tolerância centrado no objeto de consulta. Para possibilitar a execução de consultas por similaridade, aplicações que manipulam dados complexos referentes ao domínio de imagens compõem e armazenam vetores de características. Ou seja, atributos como forma, textura e cor, entre outros, são extraídos das imagens para a composição desses vetores, os quais representam o conteúdo visual das imagens por meio de valores numéricos.

O processamento de consultas por similaridade pode ser otimizado por meio de métodos de acesso métricos (MAMs). A técnica Omni é um MAM baseado no uso de representantes globais (focos) da base de dados [Traina et al. 2007], os quais são obtidos por heurísticas, e no armazenamento da distância de cada outro elemento da base aos representantes. Em uma consulta por abrangência, os representantes, as distâncias armazenadas e o elemento de consulta são usados para determinar uma região do espaço métrico denominada *mbOr* (*minimum-bounding-Omni-region*), na qual os elementos mais similares ao elemento de consulta residem, formando um conjunto de candidatos. Esses candidatos são então refinados calculando-se suas distâncias ao elemento de consulta.

3. Trabalhos Correlatos

Na literatura existem diversos trabalhos que propõem índices para ambientes de DWing e índices para o processamento de imagens, porém, não foram encontrados trabalhos que consideram esses dois aspectos conjuntamente. Portanto, os trabalhos correlatos a esse projeto abrangem: (i) DWs de imagens; (ii) índices para DWs convencionais e para o espaço métrico; e (iii) adaptações de índices bitmap para o processamento de consultas por similaridade.

Em [Arigon et al. 2007, Chen et al. 2008, Jin et al. 2010, Annibal et al. 2010], dados multimídia são incluídos em ambientes de DWing não-convencionais. Dentre esses trabalhos, a proposta mais abrangente e flexível é a descrita em [Annibal et al. 2010], a qual possui as seguintes características: (i) esquema estrela estendido para armazenar também vetores de características de imagens em tabelas de dimensão; (ii) a possibilidade de se usar diferentes camadas perceptuais, ou seja, diferentes vetores de características para representar cada imagem (ex.: um vetor para cor, outro para textura); (iii) estratégia de extração, tradução e carregamento de dados de imagens no DW estendido; (iv) e uso de consulta por abrangência para a execução de consultas OLAP baseadas em similaridade de imagens. Entretanto, esse trabalho não inclui a proposta de uma estratégia de indexação especificamente projetada para um ambiente de DWing de imagens, o que é o objetivo do ImageDW-index.

Considerando MAMs e índices para DWs convencionais, existem muitas propostas, por exemplo [Chmiel et al. 2009, Carélo et al. 2011]. Embora os índices para DWs melhorem o desempenho do processamento de consultas OLAP, eles não oferecem funcionalidades voltadas ao processamento de consultas por similaridade. Em contrapartida, embora os índices métricos melhorem o desempenho do processamento de consultas por similaridade de imagens, eles não enfocam características intrínsecas de ambientes de DWing, como a multidimensionalidade dos dados. A proposta do ImageDW-index visa considerar características de ambientes de DWing e de consultas por similaridade de imagens conjuntamente.

Por fim, os trabalhos de [Jeong and Nang 2004, Nang et al. 2010, Cha 2004] são adaptações de índices bitmap para o processamento de consultas por similaridade. Em [Jeong and Nang 2004], os vetores de características são representados por meio de bits, os quais indicam quais dimensões são representativas, ou seja, com valores relativamente maiores do que os das outras dimensões. Em [Nang et al. 2010], é proposta uma hierarquia de intervalos visando a criação de representações binárias dos vetores de características, as quais identificam as dimensões representativas entre dois objetos considerando cada intervalo. Nesses dois trabalhos, o índice bitmap é usado para o cálculo de uma distância aproximada entre um objeto de consulta e os elementos da base, gerando um conjunto de elementos candidatos a resposta, o qual é posteriormente refinado. Em [Cha 2004], realiza-se o agrupamento dos dados da base para cada dimensão dos objetos, identificando vários intervalos por dimensão. Índices bitmap são criados para cada *cluster*, indicando quais objetos pertencem aos intervalos encontrados. Dois pontos classificados em um mesmo intervalo são considerados similares naquela dimensão. Entretanto, nenhum desses trabalhos considera as especificidades de ambientes de DWing, como a grande quantidade de dados e sua organização multidimensional. Em especial, o volume de dados impacta de forma negativa nesses trabalhos, principalmente devido ao

alto custo de construção desses índices. A proposta do ImageDW-index visa enfocar esses ambientes, e suas características intrínsecas.

4. Proposta e Estágio Atual de Desenvolvimento

4.1. Descrição do ImageDW-index

Como ponto de partida, está sendo investigado o armazenamento de intervalos fixos de distâncias dos vetores de características das imagens aos representantes da base de dados, aproveitando as vantagens oferecidas pela técnica Omni. Nessa abordagem, os representantes globais possuem um conjunto fixo de intervalos de distância, cada qual com seu respectivo vetor de bits indicando a pertinência dos objetos da base aos intervalos representados. Dessa forma, a intersecção dos intervalos para formação da *mbOr* é realizada de maneira muito mais rápida.

Uma ilustração do ImageDW-index é feita na Figura 1. Em uma busca por abrangência, dada uma imagem de consulta s_q e um raio de tolerância r_q, primeiramente calcula-se a distância de s_q aos representantes (f_1, f_2, f_3) para definir os intervalos em torno de cada representante (anéis). Para f_2, por exemplo, tem-se o intervalo $[d(f_2, s_q) - r_q, d(f_2, s_q) + r_q]$. Utilizando o ImageDW-index, os objetos pertencentes à *mbOr* (interseção dos anéis de cada representante) são encontrados rapidamente por meio de operações lógicas bit a bit. A etapa de refinamento é realizada posteriormente sobre o conjunto de elementos candidatos retornados.

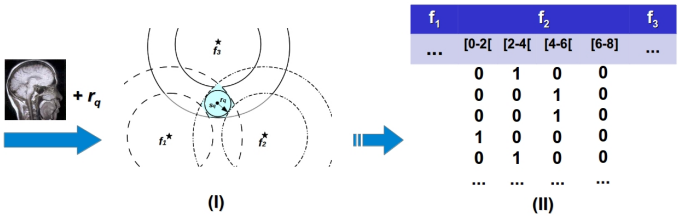

Figura 1. ImageDW-index para geração otimizada da *mbOr* (I), dada uma imagem de consulta e um raio de tolerância r_q. Cada elemento representativo (f_1, f_2, f_3) possui intervalos de distância, para os quais são gerados os índices bitmap, conforme ilustrado em (II).

4.2. Validação do ImageDW-index

Foram realizados testes de desempenho preliminares usando um conjunto de 131.656 imagens médicas, previamente processadas por cinco descritores diferentes, ou seja, foram geradas cinco diferentes camadas perceptuais para cada imagem. As imagens foram disponibilizadas pelo Grupo de Banco de Dados e Imagens (GBdi) da USP-São Carlos e os dados convencionais foram obtidos do site `www2.datasus.gov.br/datasus`. Dados sintéticos também foram gerados, utilizando o *benchmark* TPC-H [Poess and Floyd 2000]. Os testes foram realizados em uma máquina com processador AMD Phenon II Quad-Core, 4GB de memória RAM, rodando o sistema operacional Linux (Ubuntu 12.04). Ademais, foi utilizada a implementação de índices bitmap da biblioteca *open-source* FastBit [Wu 2005].

Foram definidas duas configurações para o ImageDW-index: **Conf1**, a qual considerou apenas a indexação dos intervalos de distâncias; e **Conf2**, a qual considerou a indexação dos intervalos de distâncias, bem como a de atributos convencionais usando também índice bitmap. A indexação das distâncias aos representantes globais foi feita utilizando a técnica de *binning*. Para a etapa de refinamento, foi necessário recuperar os vetores de características dos objetos na *mbOr*, dessa forma, em cada configuração foi testado se a indexação dos vetores de características foi mais vantajosa do que acessar o DW para recuperá-los.

Os testes realizados consistem da execução de várias consultas analíticas, derivadas da consulta "Quantas imagens são similares a uma dada imagem de consulta segundo um raio de abrangência de 30% nas cinco camadas perceptuais e concomitantemente são imagens geradas no hospital da macrorregião da Grande São Paulo, nos anos de 1992 e 1993, referentes a pacientes com suspeita de tumor, do estado de São Paulo e com idade entre 0 a 30 anos?" Em cada consulta, os predicados convencionais foram progressivamente eliminados de acordo com sua seletividade (do mais seletivo até a ausência total do predicado). Cada consulta foi executada 10 vezes em cada configuração e o tempo médio em segundos foi coletado.

Comparando as duas configurações, Conf2 apresentou melhores resultados em termos de tempo em segundos do que Conf1, visto que em Conf2 todos os atributos (convencionais e de imagens) são indexados, de modo que o DW é acessado apenas para recuperar os vetores de características. Ademais, em Conf1, consultas com predicados convencionais não são plenamente beneficiadas pelo índice bitmap, dessa forma, o DW deve ser acessado para filtrar os dados pelos atributos convencionais e ainda para recuperar os vetores de características para o refinamento, o que torna mais custoso o processamento dessas consultas nessa configuração.

Os mesmos testes foram realizados para comparar as configurações propostas com a estratégia de otimização de consultas definida no trabalho correlato [Annibal et al. 2010]. Conf2 apresentou melhores resultados em termos de tempo em segundos, obtendo ganhos que variaram de 20% até 62% nos testes realizados. Isso demonstrou que o ImageDW-index é capaz de prover bons resultados de desempenho no processamento de consultas OLAP que possuam predicados de similaridade de imagens. Esse bom desempenho é oriundo do uso conjunto de conceitos bem difundidos em DW, como o índice bitmap, e de conceitos bem difundidos para o armazenamento e recuperação de imagens, como a técnica Omni.

5. Considerações Finais e Próximas Atividades

Avançando no estado da arte da pesquisa em DW de imagens, é apresentado o ImageDW-index, uma estratégia de indexação voltada ao processamento eficiente de consultas analíticas envolvendo predicados de similaridade entre imagens. De acordo com o estágio atual de desenvolvimento, a abordagem une as vantagens de índices bitmap e da técnica Omni. As próximas atividades referem-se à continuidade do processo de validação da proposta atual, por meio da realização de testes de escalabilidade, submetendo as configurações propostas a grandes volumes de dados, e comparação com outros trabalhos correlatos. Pretende-se também adaptar o índice proposto em [Jeong and Nang 2004] ao ImageDW-index, de modo a criar um segundo mecanismo de filtragem.

Referências

Annibal, L., Felipe, J., Ciferri, C., and Ciferri, R. (2010). icube: A similarity-based data cube for medical images. In *CBMS*, pages 321 –326.

Arigon, A.-M., Miquel, M., and Tchounikine, A. (2007). Multimedia data warehouses: a multiversion model and a medical application. *Multimedia Tools Appl.*, pages 91–108.

Carélo, C. C. M., Pola, I. R. V., Ciferri, R. R., Traina, A. J. M., Jr, C. T., and de Aguiar Ciferri, C. D. (2011). Slicing the metric space to provide quick indexing of complex data in the main memory. *Information Systems*, pages 79 – 98.

Cha, G.-H. (2004). Efficient and flexible bitmap indexing for complex similarity queries. In Lee, Y., Li, J., Whang, K.-Y., and Lee, D., editors, *DASFAA*, pages 708–720. Springer Berlin Heidelberg.

Chaudhuri, S. and Dayal, U. (1997). An overview of data warehousing and olap technology. *SIGMOD Rec.*, pages 65–74.

Chen, M., Song, Y., Sun, Z., Chen, H., and Sang, A. (2008). Multimedia database retrieval based on data cube. In *ICALIP*, pages 1265 –1269.

Chmiel, J., Morzy, T., and Wrembel, R. (2009). Hobi: Hierarchically organized bitmap index for indexing dimensional data. In *DaWaK*, pages 87–98. Springer-Verlag.

Ciaccia, P. and Patella, M. (2002). Searching in metric spaces with user-defined and approximate distances. *TODS*, pages 398–437.

Jeong, J. and Nang, J. (2004). An efficient bitmap indexing method for similarity search in high dimensional multimedia databases. In *ICME*, pages 815 –818.

Jin, X., Han, J., Cao, L., Luo, J., Ding, B., and Lin, C. X. (2010). Visual cube and on-line analytical processing of images. In *CIKM*, pages 849–858. ACM.

Nang, J., Park, J., Yang, J., and Kim, S. (2010). A hierarchical bitmap indexing method for similarity search in high-dimensional multimedia databases. *JISE*, pages 393–407.

O'Neil, P. and Graefe, G. (1995). Multi-table joins through bitmapped join indices. *SIGMOD Rec.*, pages 8–11.

O'Neil, P. and Quass, D. (1997). Improved query performance with variant indexes. In *SIGMOD Rec.*, pages 38–49. ACM.

Poess, M. and Floyd, C. (2000). New tpc benchmarks for decision support and web commerce. *SIGMOD Rec.*, pages 64–71.

Traina, Jr., C., Filho, R. F., Traina, A. J., Vieira, M. R., and Faloutsos, C. (2007). The omni-family of all-purpose access methods: a simple and effective way to make similarity search more efficient. *VLDB*, pages 483–505.

Wu, K. (2005). Fastbit: an efficient indexing technology for accelerating data-intensive science. *JPCS*, page 556.

Wu, K., Stockinger, K., and Shoshani, A. (2008). Breaking the curse of cardinality on bitmap indexes. In *SSDBM*, pages 348–365. Springer-Verlag.