UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DAINF - DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

CRISTIANO JOSÉ MENDES MATSUI

CONSULTAS POR SIMILARIDADE EM BASES DE DADOS COMPLEXAS UTILIZANDO TÉCNICAS OMNI EM SGBDR

TRABALHO DE CONCLUSÃO DE CURSO

PATO BRANCO
2017

CRISTIANO JOSÉ MENDES MATSUI

CONSULTAS POR SIMILARIDADE EM BASES DE DADOS COMPLEXAS UTILIZANDO TÉCNICAS OMNI EM SGBDR

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia de Computação da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Dr. Ives Renê Venturini Pola

Coorientadora: Dra. Fernanda Paula Barbosa Pola

PATO BRANCO
2017

Altere este texto inserindo a dedicatória do seu trabalho.

AGRADECIMENTOS

Edite e coloque aqui os agradecimentos às pessoas e/ou instituições que contribuíram para a realização do trabalho.

É obrigatório o agradecimento às instituições de fomento à pesquisa que financiaram total ou parcialmente o trabalho, inclusive no que diz respeito à concessão de bolsas.

Eu denomino meu campo de Gestão do Conhecimento, mas você não pode gerenciar conhecimento. Ninguém pode. O que pode fazer - o que a empresa pode fazer - é gerenciar o ambiente que otimize o conhecimento. (PRUSAK, Laurence, 1997).

RESUMO

MATSUI, Cristiano. Consultas por similaridade em bases de dados complexas utilizando técnicas OMNI em SGBDR. 2017. 11 f. Trabalho de Conclusão de Curso – Curso de Engenharia de Computação, Universidade Tecnológica Federal do Paraná. Pato Branco, 2017.

O Resumo é um elemento obrigatório em tese, dissertação, monografia e TCC, constituído de uma seqüência de frases concisas e objetivas, fornecendo uma visão rápida e clara do conteúdo do estudo. O texto deverá conter no máximo 500 palavras e ser antecedido pela referência do estudo. Também, não deve conter citações. O resumo deve ser redigido em parágrafo único, espaçamento simples e seguido das palavras representativas do conteúdo do estudo, isto é, palavras-chave, em número de três a cinco, separadas entre si por ponto e finalizadas também por ponto. Usar o verbo na terceira pessoa do singular, com linguagem impessoal, bem como fazer uso, preferencialmente, da voz ativa. Texto contendo um único parágrafo.

Palavras-chave: Palavra. Segunda Palavra. Outra palavra.

ABSTRACT

MATSUI, Cristiano. Similarity queries in complex databases using OMNI techniques in RDBMS. 2017. 11 f. Trabalho de Conclusão de Curso – Curso de Engenharia de Computação, Universidade Tecnológica Federal do Paraná. Pato Branco, 2017.

Elemento obrigatório em tese, dissertação, monografia e TCC. É a versão do resumo em português para o idioma de divulgação internacional. Deve ser antecedido pela referência do estudo. Deve aparecer em folha distinta do resumo em língua portuguesa e seguido das palavras representativas do conteúdo do estudo, isto é, das palavras-chave. Sugere-se a elaboração do resumo (Abstract) e das palavras-chave (Keywords) em inglês; para resumos em outras línguas, que não o inglês, consultar o departamento / curso de origem.

Keywords: Word. Second Word. Another word.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
DECOM	Departamento de Computação

LISTA DE SÍMBOLOS

Γ	Letra grega Gama
λ	Comprimento de onda
$\in$	Pertence

LISTA DE ALGORITMOS

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo será apresentado uma contextualização do problema, assim como o estado da arte abordado por este trabalho. Também será introduzido um tipo de consulta não-nativo ao SGBDR, a consulta por similaridade, assim como os tipos de consultas por similaridade que serão utilizados neste trabalho como a consulta por abrangência e a consulta por k-vizinhos mais próximos. Finalmente, será apresentado a organização deste documento.

1.1 CONSIDERAÇÕES INICIAIS

Nos anos recentes, foi notado um grande aumento no tráfego e armazenamento de diferentes aplicações e dados multimídias, como imagens, áudio, vídeo, impressões digitais, séries temporais, sequências de proteínas, etc. Estes tipos de dados, que apresentam muito mais atributos do que simples numerais ou pequenas cadeias de caracteres, são conhecidos como dados complexos (ZIGHED et al., 2008).

Quando tratados por um Sistema Gerenciador de Banco de Dados Relacional (SGBDR), não suportam comparações com os operadores conhecidos como "big six" da linguagem SQL: $=$, $\neq$, $<$, $>$, $\leq$, $\geq$. Esse fato limita muito as comparações entre dados complexos inseridos em um SGBDR, ocasionando um grande problema no contexto de base de dados, uma vez que os principais sistemas de gerenciamento de base de dados são relacionais (DB-ENGINES, 2017). Com isso, tornou-se necessária a concepção de novos tipos de comparadores, como buscas por similaridade.

Estas consultas por similaridade se aplicam de maneira geral a muitos dos tipos de dados complexos (BARIONI et al., 2009). Embora equiparar duas imagens médicas (como tomografias de pacientes distintos) raramente produza um resultado diferente de falso, procurar por imagens semelhantes à original faz mais sentido e retorna resultados mais relevantes.

1.2 TIPOS DE CONSULTAS

Dentre os operadores de consulta por similaridade os mais comuns são as consultas por abrangência (*range query: Rq*) e consulta aos k-vizinhos mais próximos (*k-nearest neighbor querry: kNNq*).

As consultas por abrangência *Rq*(s_q, ξ) recebem como parâmetro um elemento s_q do domínio de dados (elemento central da consulta) e um limite máximo de dissimilaridade ξ. O resultado é o conjunto de elementos da base que diferem do elemento central da consulta por no máximo a dissimilaridade indicada.

Definição 1 *Seja S um atributo complexo de um domínio $\mathbb{S}$ sobre o qual a condição de similaridade é expressada, seja d uma função de distância, seja ξ o limiar de dissimilaridade e*

seja $s_q \in \mathbb{S}$ o elemento central de consulta. A consulta Rq(s_q , ξ) retorna todos os elementos $s_i \in \mathbb{S}$ que possuem o valor do atributo S distantes até um máximo de ξ deste atributo referente ao elemento central da consulta:

$$Rq(s_q, \xi) : S = \{s_i \in S \mid d(s_i(S), s_q) \leq \xi\} \tag{1}$$

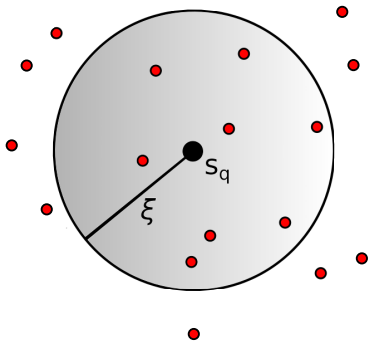


Figura 1 – Exemplo de consulta por abrangência

Fonte: Autoria Própria

A figura 1 exemplifica este tipo de consulta. Tomando s_q e ξ como parâmetros da consulta, todos os elementos contidos pelo raio de abrangência fazem parte do conjunto resposta da consulta. s_q não necessariamente precisa pertencer ao conjunto de dados de pesquisa, devendo apenas pertencer ao mesmo domínio de dados.

Uma consulta aos k-vizinhos mais próximos *kNNq(s_q, k)* também recebe como um de seus parâmetros um elemento central da consulta s_q, e um número inteiro k de vizinhos desejados, e retorna como resultado o conjunto dos k elementos com a menor dissimilaridade em relação ao elemento central da consulta s_q (POLA, 2010).

Definição 2 *Seja S um atributo complexo de um domínio $\mathbb{S}$ sobre o qual a condição de similaridade é expressada, seja d uma função de distância, seja $k \in \mathbb{N}^*$ a quantidade de elementos desejados e seja $s_q \in \mathbb{S}$ o elemento central de consulta. A consulta kNNq(s_q , k) retorna k elementos $s_i \in \mathbb{S}$ que possuem o valor do atributo S menos distantes do valor deste atributo referente ao elemento central da consulta(FERREIRA et al., 2009):*

$$kNNq(s_q, k) : S = \{s_i \in S \mid \forall\ s_j \in S - S^{'}, d(s_i, s_q) \leq d(s_j, s_q)\}, \tag{2}$$

onde $S^{'} = 0$, se $i = 1$ e $S^{'} = \{s_1, \ldots\ldots, s_{(i-1)}\}$, se $1 < i \leq k$.

A figura 2 exemplifica este tipo de operação. Nesta consulta, os parâmetros foram o elemento central da consulta s_q e o número de elementos k a serem encontrados igual a 4. Os elementos ligados ao elemento central foram os mais próximos a este, portanto apenas eles fazem parte do conjunto resposta da consulta.

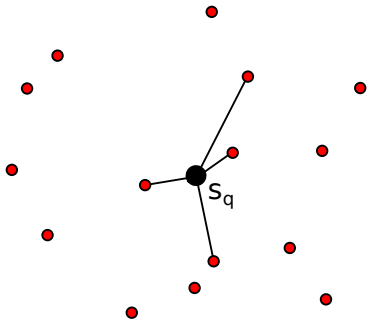


Figura 2 – Exemplo de consulta por k-vizinhos mais próximos

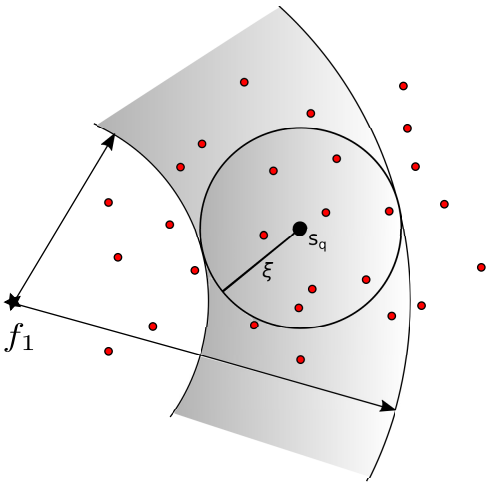
Fonte: Autoria Própria

O SGBDR não possui suporte nativo a estes tipos de consulta, mas é possível construir estas consultas utilizando ferramentas existentes em um banco de dados relacional (como a *B-tree*).

A solução abordada por esta proposta é a do uso de técnicas Omni, presentes no trabalho de (FILHO et al., 2001). Um número calculado de elementos do conjunto de dados são selecionados como "focos", e utilizados para podar cálculos desnecessários de distâncias, fazendo uso da desigualdade triangular.

A base da técnica Omni é calcular previamente as distâncias de todos os elementos para todos os focos selecionados, armazenando estas distâncias no banco. Quando uma consulta por similaridade (como uma consulta por abrangência) é realizada, são conhecidas as distâncias entre o elemento central da consulta s_q e o raio de abrangência ξ. Considerando um foco f_1 e utilizando a desigualdade triangular, elementos que possuem uma distância entre o foco escolhido menor do que a distância de s_q até o foco menos o valor ξ serão descartados do conjunto de elementos necessários para os cálculos de distância com o elemento central. Simetricamente, elementos cuja distância até o foco seja maior do que a distância de s_q até o foco mais o valor do raio de abrangência ξ também serão descartados. Com isso, ocorre uma grande redução do número de cálculos necessários para fornecer o conjunto resposta. Essa poda também pode ser realizada por mais de um foco.

Figura 3 – Consulta por abrangência pela técnica Omni utilizando um foco



Fonte: Autoria Própria

A figura 3 ilustra a poda no número de cálculos. Apenas os elementos na área sombreada terão as suas distâncias em relação ao centro da consulta calculadas, pois estão no conjunto de elementos que não foram descartados utilizando a desigualdade triangular com as distâncias previamente calculadas em relação ao foco. O armazenamento das distâncias de cada foco f_i para cada outro elemento s_k será feito utilizando uma estrutura de indexação que implementa os conceitos da técnica Omni com a estrutura da B$^+$-tree, originando uma nova estrutura chamada de Omni-Btree. As distâncias serão armazenadas em l Omni-Btrees, sendo l o número de focos criados para a base de dados (FILHO et al., 2001).

1.3 OBJETIVOS

Os objetivos encontram-se divididos em objetivos gerais, referentes ao resultado obtido com a conclusão deste trabalho e objetivos específicos, ilustrando etapas intermediárias necessárias para alcançar o objetivo geral.

1.3.1 OBJETIVOS GERAIS

O principal foco deste trabalho é a construção de um sistema de consultas por similaridade em uma base de imagens utilizando técnicas da família Omni para reduzir o custo computacional das operações de consulta.

1.3.2 OBJETIVOS ESPECÍFICOS

- Modelar o banco de dados para atender a problemática apresentada;
- Aplicar os extratores de características das imagens utilizadas;
- Povoar o banco de dados com as imagens e os valores de suas características;
- Criar a estrutura Omni necessária para a filtragem dos cálculos;

- Analisar e comparar os resultados obtidos.

1.4 ORGANIZAÇÃO DO TRABALHO

A estrutura deste trabalho apresenta mais detalhes sobre os tipos de consulta por similaridade utilizados, assim como os seus algoritmos computacionais. Posteriormente, explana os conceitos e técnicas da família Omni utilizada para melhorar a performance das consultas. Também realiza a abordagem sobre os extratores de características utilizados para a base de imagens utilizada neste trabalho.

2 CONCEITOS

O presente capítulo abordará os conceitos necessários para uma melhor compreensão dos artifícios utilizados para tratar o problema apresentado no Capítulo 1, assim como uma explanação sobre os extratores de características de imagens que serão utilizados e o estado da arte atual.

2.1 ESPAÇOS MÉTRICOS

Uma métrica em um conjunto $\mathbb{S}$ é uma função $d : \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}^+$ que associa a cada par ordenado de elementos (s_1, s_2) um número real $d(s_1,s_2)$ chamado de distância de s_1 a s_2 (LIMA, 1977).

Um espaço métrico $\mathbb{M}$ é um par $<\mathbb{S}, d>$ onde $\mathbb{S}$ é um conjunto de elementos e d é uma métrica (ou função de distância). Esta distância $d(s_1,s_2)$ pode ser compreendido como uma medida de dissimilaridade entre dois elementos. Quanto menor esta distância entre dois elementos, mais semelhantes eles são.

Definição 3 *Seja $\mathbb{S}$ um conjunto não-vazio de elementos e $d(s_1,s_2)$ uma métrica definida sobre $\mathbb{S} \times \mathbb{S}$. O par $<\mathbb{S}, d>$ é chamado de espaço métrico desde que d satisfaça as seguintes propriedades:*

1. *$d(s_1,s_1) = 0$ (identidade);*
2. *Se $s_1 \neq s_2$ então $d(s_1,s_2) > 0$ (não-negatividade);*
3. *$d(s_1,s_2) = d(s_2,s_1)$ (simetria);*
4. *$d(s_1,s_3) \leq d(s_1,s_2) + d(s_2,s_3)$ (desigualdade triangular)*

onde s_1, s_2, $s_3 \in \mathbb{S}$.

A importância prática do espaço métrico para este trabalho reside fortemente na quarta propriedade apresentada pela Definição 3. A desigualdade triangular é fundamentada na geometria euclidiana, onde a soma do comprimento de dois lados de um triângulo não pode ser superior ao comprimento do terceiro lado, como ilustra a figura 4.

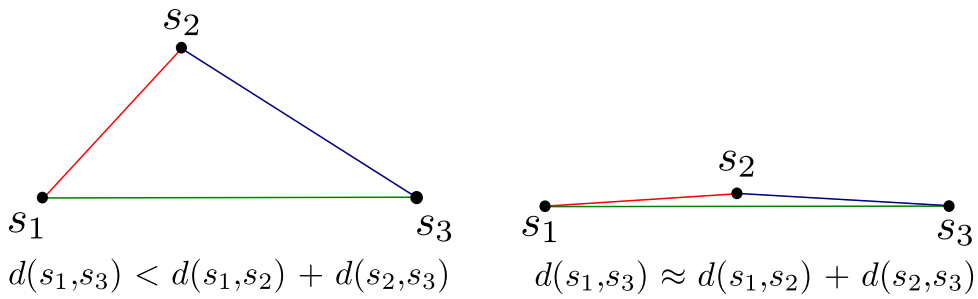


Figura 4 – Exemplo geométrico da desigualdade triangular

Fonte: Autoria Própria

O emprego da desigualdade triangular neste trabalho será detalhado no capítulo **??**.

2.2 FUNÇÕES DE DISTÂNCIA

Para verificar a similaridade entre dois elementos de um domínio, é utilizada uma função de distância. Esta função recebe como parâmetro um par de elementos do conjunto e retorna o valor da dissimilaridade entre eles. Quanto mais próximo de zero, mais similares os elementos são.

A importância do uso de uma função de distância para este trabalho é fornecer uma métrica de comparação entre os elementos. Além de necessária para a consulta por similaridade, a função de distância também é utilizada para evitar cálculos desnecessários, fornecendo uma métrica empregada pela técnica Omni.

Dentre as funções de distância, será abordada a metrica Minkowski. Esta métrica é a mais utilizada para cálculos de índice de similaridade, e tem como resultado o valor da dissimilaridade entre dois elementos (JAIN; DUBES, 1988).

Definição 4 *Sejam* $s_1 = \{s_{11},s_{12},...,s_{1n}\}$ *e* $s_2 = \{s_{21},s_{22},...,s_{2n}\}$ *dois vetores de dimensionalidade* n *pertencentes ao conjunto de elementos* $\mathbb{S}$*, a distância Minkowski entre esses dois elementos é dada por:*

$$d(s_1,s_2) = \sqrt[p]{\sum_{i=1}^{n} |s_{1i} - s_{2i}|^p} \tag{3}$$

Para o caso em que $p = 2$ (L_2), a distância Minkowski torna-se a tradicional distância euclidiana, amplamente utilizada para distância entre vetores.

2.3 ESTRUTURAS DE INDEXAÇÃO

Dados complexos costumam apresentar tamanho físico muito elevado quando comparados com dados numéricos ou pequenas cadeias de caracteres, que são os tipos de dados mais comuns em bancos de dados tradicionais. Para responder consultas que envolvam dados complexos, são necessários mais acessos a disco em relação a uma consulta que envolva tipos de dados mais simples, como os previamente mencionados. Estes acessos são custosos e precisam ser reduzidos para um melhor desempenho do banco.

Uma maneira de tornar o acesso a disco mais eficiente é evitar movimentar grandes porções do banco de dados do disco para a memória, fazendo o uso de índices dentro do banco de dados. Um índice em um SGBDR funciona de maneira semelhante a um índice de um livro. Para procurar um tópico específico, é possível consultar o índice no fim do livro e descobrir qual o número da página correspondente ao tópico, contornando a necessidade da leitura sequencial do livro até o tópico procurado. Os índices são armazenados em ordem e apresentam um tamanho muito menor do que um capítulo do livro, reduzindo o esforço necessário para a sua consulta (SILBERSCHATZ; KORTH; SUDARSHAN, 2011).

2.4 EXTRATORES DE CARACTERÍSTICAS

O principal foco deste trabalho é o emprego destas técnicas para bases de dados constituídas por imagens. Para isto, torna-se necessário o uso de uma miríade de extratores de características das imagens, para um maior refinamento do uso de consultas por similaridade. Estas características podem se referir a: atributos visuais (cor, forma, textura), atributos lógicos (identificação de elementos) e atributos semânticos (identificação de emoções humanas).

As características visuais podem ser utilizadas como histogramas de cores para a análise de cor, matrizes de co-ocorrência para a análise de textura e métodos baseados em contorno para a análise de forma. Geralmente, consultas são feitas utilizando uma combinação destas características, e não apenas uma delas.

Dado uma palheta discreta de cores definida por alguns eixos de cor, o histograma de cores é obtido através da discretização das cores da imagem e contagem do número de vezes que cada cor discreta ocorre na matriz da imagem (SWAIN; BALLARD, 1991). As vantagens do uso do histograma de cores é a sua simplicidade computacional e pouca sensibilidade a alterações na imagem (rotação e translação), particularmente útil para a representação de objetos tridimensionais. Entretanto, duas imagens completamente diferentes podem apresentar o mesmo histograma de cores.

Para a análise de textura, o objetivo e conseguir distinguir regiões que apresentam cores similares (como folhagem e grama), analisando o padrão de variação dessas cores. A técnica mais utilizada analisa conjunto de pares de pixels da imagem em tons de cinza e monta estruturas com informações características. A principal estrutura utilizada nesta técnica é a "Matriz de co-ocorrência".

Diversas medidas que podem ser extraídas destas matrizes de co-ocorrência estão presentes no trabalho de (HARALICK; SHANMUGAM; DINSTEIN, 1973). Os extratores de características das imagens utilizados neste trabalho são do framework Arboretum, desenvolvido pelo Grupo de Bases de Dados e Imagens (GBDI) da Universidade de São Paulo - campus São Carlos.

3 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Cada capítulo deve conter uma pequena introdução (tipicamente, um ou dois parágrafos) que deve deixar claro o objetivo e o que será discutido no capítulo, bem como a organização do capítulo.

4 CONCLUSÃO

Parte final do texto, na qual se apresentam as conclusões do trabalho acadêmico. É importante fazer uma análise crítica do trabalho, destacando os principais resultados e as contribuições do trabalho para a área de pesquisa.

4.1 TRABALHOS FUTUROS

Também deve indicar, se possível e/ou conveniente, como o trabalho pode ser estendido ou aprimorado.

4.2 CONSIDERAÇÕES FINAIS

Encerramento do trabalho acadêmico.

Referências

BARIONI, M. C. N. et al. Seamlessly integrating similarity queries in sql. **Software: Practice and Experience**, John Wiley & Sons, Ltd., v. 39, n. 4, p. 355–384, 2009. ISSN 1097-024X. Disponível em: <http://dx.doi.org/10.1002/spe.898>. Citado na página 1.

DB-ENGINES. **DB-Engines Ranking**. 2017. Disponível em: <https://db-engines.com/en/ranking>. Acesso em: 30 de agosto de 2017. Citado na página 1.

FERREIRA, M. R. P. et al. Identifying algebraic properties to support optimization of unary similarity queries. v. 450, 05 2009. Citado na página 2.

FILHO, R. F. S. et al. Similarity search without tears: The omni family of all-purpose access methods. In: **Proceedings of the 17th International Conference on Data Engineering**. Washington, DC, USA: IEEE Computer Society, 2001. p. 623–630. ISBN 0-7695-1001-9. Disponível em: <http://dl.acm.org/citation.cfm?id=645484.656543>. Citado 2 vezes nas páginas 3 e 4.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. Textural features for image classification. **IEEE Transactions on Systems, Man, and Cybernetics**, SMC-3, n. 6, p. 610–621, Nov 1973. ISSN 0018-9472. Citado na página 8.

JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988. ISBN 0-13-022278-X. Citado na página 7.

LIMA, E. **Espaços métricos**. [S.l.]: Instituto de Matemática Pura e Aplicada, CNPq, 1977. (Projeto Euclides). Citado na página 6.

POLA, I. R. V. **Explorando conceitos da teoria de espaços métricos em consultas por similaridade sobre dados complexos**. Agosto 2010. Tese (Doutorado em Ciência da Computação) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2010. Citado na página 2.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Database system concepts**. 6. ed. [S.l.]: McGraw-Hill, 2011. Citado na página 7.

SWAIN, M. J.; BALLARD, D. H. Color indexing. **International Journal of Computer Vision**, v. 7, n. 1, p. 11–32, Nov 1991. ISSN 1573-1405. Disponível em: <https://doi.org/10.1007/BF00130487>. Citado na página 8.

ZIGHED, D. A. et al. **Mining Complex Data**. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2008. ISBN 3540880666, 9783540880660. Citado na página 1.