Recomendação de Consultas de Banco de Dados utilizando Agrupamentos de Usuários

Márcio de Carvalho Saraiva, Carlos Eduardo Santos Pires (Orientador), Leandro Balby Marinho (Orientador)

Pós-Graduação em Ciência da Computação
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande (UFCG)
Campina Grande, PB, Brasil

marcio@copin.ufcg.edu.br, {cesp,lbmarinho}@dsc.ufcg.edu.br

Nível: Mestrado
Ano de ingresso no programa: 2012
Exame de qualificação: Dezembro de 2012
Época esperada de conclusão: Fevereiro de 2014

Abstract. *Database systems are becoming increasingly popular in the scientific community to support data exploration. In this scenario, users may not have the necessary knowledge about the domain of the database or how to formulate SQL queries to analyze the data. To solve this problem, we developed a study on the most recent techniques for query recommendation to improve them in such a manner that the user can receive better recommendations. In this proposal we discuss the main challenges of query recommendation systems and how clustering users can improve recommendations. Preliminary experimental results using real user query logs show that our study can generate effective query recommendations.*

Keywords: Query Recommendation, SQL, Databases, Query logs, Clustering.

1. Introdução

A humanidade produz milhares de exabytes (bilhões de gigabytes de dados) por ano (The Economist, 2010), o que faz com que sistemas de bancos de dados se tornem cada vez mais presentes na vida de usuários e desenvolvedores. Neste sentido, diversas empresas passaram a disponibilizar seus dados para estudos e pesquisas. Por exemplo, o site do projeto *UCSC Genome Browser* (Genome, 2013) fornece acesso a um banco de dados sobre genética, e o site do projeto *SkyServer* (*Skyserver*, 2013) disponibiliza um banco de dados sobre astronomia. O acesso a essas bases de dados ocorre por meio de aplicações que possuem uma interface web e permitem que vários usuários espalhados pelo mundo enviem consultas, normalmente escritas na linguagem SQL. Em geral, os bancos de dados apresentam esquemas complexos contendo uma grande quantidade de elementos (e.g. tabelas e colunas). Por serem ambientes abertos, os sites que disponibilizam seus bancos de dados na web possuem vários tipos de usuário. Entre esses tipos de usuário, encontramos aqueles que ainda não estão familiarizados com os elementos do esquema do banco de dados e podem não saber formular consultas que poderiam recuperar dados relevantes para suas necessidades.

Neste contexto, diversos trabalhos têm procurado identificar características em consultas de bancos de dados realizadas por usuários e a consequência da exploração dessas características para a recomendação de consultas. Julien Aligon, et al. (2013) e Marcel e Negre (2011) revisaram a literatura sobre técnicas de recomendação em banco de dados e observaram que faltam concepções mais refinadas dos usuários e de seus papéis com relação às consultas realizadas. Essa concepção auxiliaria na produção de técnicas e ferramentas que atendessem melhor aos anseios de perfis específicos de usuários, o que faz com que as atividades do usuário que possuam interação com o banco de dados sejam mais proveitosas, retornando dados que são importantes para suas consultas. Diferentemente, verificamos que não há conhecimento sobre técnicas para agrupamento de usuários do banco de dados em perfis que utilizam padrões de comportamento de usuários considerando características extraídas do histórico de consultas realizadas pelos mesmos. O agrupamento de usuários pode auxiliar na etapa de seleção das consultas que serão utilizadas nas técnicas de recomendação.

Neste trabalho, propomos uma técnica de mineração de padrões de acesso a banco de dados para recomendações de consultas. Essa técnica é baseada em agrupamentos de usuários construídos segundo características extraídas do histórico de consultas dos mesmos. A proposta visa auxiliar os usuários não familiarizados com o esquema de um banco de dados e que, por isso, têm dificuldade em formular consultas que poderiam recuperar dados relevantes.

2. Trabalhos relacionados

Foi realizado um levantamento do estado da arte, referente à área de recomendação de consultas de banco de dados. Com isso, observou-se que existem trabalhos que utilizam consultas realizadas por usuários em sistemas de banco de dados para auxiliar técnicas de recomendação de consultas SQL. Recentemente, Marcel e Negre (2011) afirmaram que há apenas as tentativas de Chatzopoulou et al. (2011) e Stefanidis et al., (2009) de formalizar recomendações de consultas de banco de dados para a exploração de esquema de banco de dados. Stefanidis et al. (2009) propõem uma matriz relacionando

usuários x consultas, com o intuito de medir a utilidade de uma consulta para um usuário. Esta utilidade é igual ao número de vezes que o usuário realizou a consulta. A técnica proposta por Chatzopoulou et al. (2011) de recomendação de consultas para exploração interativa de bancos de dados se enquadra nesta abordagem, embora as entradas da técnica sejam diferentes (sessões de acesso x tuplas recuperadas).

Khoussainova et al. (2010) e Akbarnejad et al. (2010) focam em fragmentos (atributos, tabelas, junções e predicados) de consultas realizadas para recomendações de novas consultas e consideram, portanto, as sessões de acesso dos usuários x matriz de fragmentos de consulta. A partir de um fragmento de consulta, o sistema procura outros fragmentos do histórico de consultas dos usuários que são similares entre si que compõem um conjunto Q'. O trabalho de Akbarnejad et al. (2010) realiza recomendações de consultas baseadas no histórico de consultas contendo fragmentos de Q'. A técnica proposta por Khoussainova et al. recomenda os fragmentos mais prováveis de Q' sabendo-se que o fragmento inicial é Q.

Buscando identificar características dos usuários de banco de dados, o trabalho de Lyes Limam et al. (2010) extrai o interesse (partes mais visualizadas do esquema e dos dados) dos usuários para expandir consultas, observando o histórico de consultas dos mesmos, sem a necessidade do preenchimento de formulários de preferências e interesses. O trabalho de Zhang e Nasraoui (2006) propõe uma técnica de mineração de dados composta por duas etapas: a primeira extrai o comportamento sequencial dos usuários (ou seja, a ordem de tabelas ou resultados que o usuário acessa) e a segunda analisa a similaridade entre consultas realizadas pelos usuários.

Há também trabalhos que expandem consultas SQL realizadas pelos usuários utilizando registros de preferências de cada usuário, como as pesquisas de Yang, Procopiuc e Srivastava (2009), e Koutrika e Ioannidis (2004), buscando adequá-las ao perfil dos usuários em sistemas de banco de dados; e trabalhos como o de Stefanidis et al. (2009) que acrescenta mais dados aos resultados retornados pelas consultas por meio da realização implícita de consultas semelhantes.

Todos esses trabalhos trouxeram contribuições significativas para a área, porém não foram encontrados em nossa revisão bibliográfica trabalhos que estudam padrões de comportamento de usuários considerando características extraídas do histórico de consultas realizadas pelos mesmos. Esse conhecimento pode auxiliar no agrupamento de usuários do banco de dados, o que facilita a escolha mais precisa de históricos de consultas de usuários para realizar recomendações e, assim, aprimorar métricas de avaliação da qualidade de retorno de técnicas de recomendação de consultas de banco de dados.

3. Estado atual do trabalho

Até o presente momento, foram selecionadas a base de dados que está sendo utilizada em nossos estudos e as técnicas de recomendação que servirão de *baseline* para a pesquisa. Além disso, foram feitos testes para avaliar a qualidade de retorno dessas técnicas, para isso foram empregadas métricas comumente utilizadas na literatura, *precision* (redefinida neste trabalho como a fração entre a quantidade de consultas retornadas como recomendação que são relevantes e a quantidade total de consultas que foram retornadas como recomendação) e *recall* (fração entre a quantidade de consultas

retornadas como recomendação que são relevantes e a quantidade total de consultas relevantes) de três técnicas selecionadas.

3.1. Base de dados selecionada

Utilizamos em nosso estudo uma base de dados disponibilizada pelo site *Skyserver*, constituída pelo histórico de consultas dos usuários do sistema durante o período de 03/2010 à 07/2011. Na base podemos observar o histórico de consultas de 447 usuários, com média aproximada de 387 consultas por usuário, constituindo 79.538 consultas executadas.

3.2. Técnicas selecionadas

Primeiramente agrupamos as técnicas estudadas de acordo com o tipo de entrada utilizado, apresentados na Tabela 1.

Tabela 1. Agrupamento das técnicas estudadas de acordo com o tipo de entrada utilizado

Usuários x Consultas	Usuários x Tuplas Recuperadas	Usuários x Fragmentos de consultas	Usuários x Esquema e dados	Usuários e registros de preferências do usuário
Stefanidis et al. (2009)	Chatzopoulou et al. (2011)	Akbarnejad et al. (2010)	Lyes Limam et al. (2010)	Yang, Procopiuc e Srivastava (2009)
		Khoussainova et al. (2010)	Zhang e Nasraoui (2006)	Koutrika e Ioannidis (2004)

Os trabalhos que utilizam usuários e registros de preferências do usuário necessitam de uma fase de obtenção de preferências do usuário, seja de forma direta com o preenchimento de formulários ou indireta com a mineração de padrões. No entanto, quando trabalhamos com um esquema de banco de dados grande[1], comum em sistemas nos quais o banco de dados está disponível na internet para qualquer usuário realizar consultas, essa fase de obtenção de preferências se torna inadequada. Por esse motivo, as técnicas pertencentes a este grupo não foram selecionadas para estudo.

As técnicas que utilizam usuários x esquemas e dados são aconselhadas para uso em ambientes em que os dados e o esquema do banco de dados pouco são alterados com o passar do tempo. Essas técnicas são sensíveis a alterações, pois levam em consideração os elementos do esquema e podem proporcionar recomendações diferentes para a mesma entrada se o esquema utilizado for alterado. Como a base de dados utilizada em nossos estudos é sujeita a alteração de dados recorrentes e o esquema de dados é grande, esse tipo de técnica de recomendação não é aconselhado.

Assim, foram selecionadas três tipos de técnicas de recomendação de consultas, usuários x consultas, usuários x tuplas recuperadas, usuários x matriz de fragmentos de consultas, que utilizam como parâmetros de entrada apenas o histórico de consultas dos usuários. Desses tipos de técnica de recomendação, foram selecionadas as técnicas

[1] Por exemplo, o esquema do banco de dados do projeto *SkyServer* contém 91 tabelas.

citadas nos trabalhos mais recentes para serem utilizadas em nossos estudos. São eles: Akbarnejad et. al. (2010), Stefanidis et. al. (2009) e Chatzopoulou et. al. (2011).

A base de dados utilizada em nossa pesquisa possui diversos tipos de retorno, como imagens e dados numéricos que podem sofrer alteração com o tempo, assim, desenvolvemos a técnica proposta por Chatzopoulou et. al. (2011) substituindo as tuplas que seriam utilizadas no algoritmo por consultas dos usuários. Assim, adaptamos em particular este trabalho para entrada de usuários x consulta, tornando-a mais adequada para nossos estudos utilizando a base de dados do *Skyserver* e mais semelhante ao tipo de entrada utilizada nas outras técnicas selecionadas para serem estudadas.

3.3 Análise das técnicas

Foi realizada uma análise das recomendações de consultas após a aplicação das técnicas selecionadas para 100 usuários selecionados aleatoriamente dentre os 477 usuários do histórico de consultas do banco de dados do site *SkyServer* em duas etapas. Na primeira etapa da análise, calculamos duas métricas de avaliação da qualidade dos retornos das técnicas de recomendação estudadas. Na segunda etapa, comparamos as técnicas entre si por meio de testes estatísticos para definirmos qual será o *baseline* para nossa pesquisa.

Na primeira etapa, o histórico de consultas de cada usuário foi dividido pela metade. A primeira parte foi utilizada como entrada para o algoritmo e a segunda parte utilizada como teste. Se as consultas retornadas como recomendações pelas técnicas estivessem presentes na parte de teste, essas seriam consideradas consultas relevantes para o cálculo das duas métricas estudadas. Ao fim desta etapa pudemos verificar que a média de *precision* e *recall* das três técnicas é baixa (~20%). A Figura 1 ilustra uma amostra dos experimentos executados para medir as métricas estudadas utilizando três históricos de usuários selecionados aleatoriamente.

	Técnicas/ Precision		
Usuarios	Chatzopoulou	Akbarnejad	Stefanidis
s0,s1,s2	1,00	0,07	0,18
s3,s4,s5	0,01	0,00	0,00
s6,s7,s8	0,00	0,00	0,00
s1,s2,s3	0,43	0,02	0,18
s2,s3,s4	0,60	0,90	0,87
s4,s5,s6	0,78	0,00	0,39
s5,s6,s7	1,00	0,71	0,36
s7,s8,s0	0,50	0,60	0,02
s8,s0,s1	0,25	0,15	0,34
s0,s1,s3	0,67	0,08	0,17
s1,s3,s5	0,43	0,02	0,17
s2,s4,s6	0,60	0,90	0,87
s3,s5,s7	0,01	0,00	0,00
s4,s6,s8	0,48	0,00	0,00
s5,s7,s9	0,17	0,71	0,00
s7,s8,s9	1,00	0,40	0,02
s9,s8,s6	0,40	0,10	0,38
s8,s7,s5	0,25	0,15	0,33
s7,s6,s4	0,50	0,60	0,08
s6,s5,s3	0,00	0,00	0,00
média	0,45	0,27	0,22

	Técnicas/ Recall		
Usuarios	Chatzopoulou	Akbarnejad	Stefanidis
s0,s1,s2	0,03	0,10	0,25
s3,s4,s5	0,00	0,00	0,00
s6,s7,s8	0,00	0,00	0,00
s1,s2,s3	0,10	0,06	0,58
s2,s3,s4	0,02	0,61	0,59
s4,s5,s6	0,46	0,00	0,46
s5,s6,s7	0,18	0,13	0,07
s7,s8,s0	0,01	0,02	0,02
s8,s0,s1	0,02	0,16	0,02
s0,s1,s3	0,06	0,11	0,24
s1,s3,s5	0,10	0,06	0,55
s2,s4,s6	0,02	0,61	0,59
s3,s5,s7	0,00	0,00	0,00
s4,s6,s8	0,12	0,00	0,00
s5,s7,s9	0,03	0,13	0,00
s7,s8,s9	0,02	0,02	0,02
s9,s8,s6	0,06	0,15	0,57
s8,s7,s5	0,02	0,16	0,35
s7,s6,s4	0,01	0,02	0,02
s6,s5,s3	0,00	0,00	0,00
média	0,06	0,12	0,22

Figura 1. Amostra de *precision* e *recall*, observando os retornos das técnicas selecionadas utilizando como entrada três usuários.

Estudamos a distribuição dos valores obtidos com aplicação das métricas em cada uma das técnicas e visualizamos que as distribuições não são do tipo normal.

Assim, por meio do teste de Kruskal-Wallis, verificamos que todas as técnicas são equivalentes analisando *precision* e *recall*. Por esse motivo, todas elas serão consideradas em nossos estudos como *baseline* da pesquisa. Verificamos também que, quanto maior o número de históricos de consultas de usuários que são utilizados como entrada para as técnicas de recomendação estudadas, menores serão as médias encontradas para as métricas observadas.

Assim, pretendemos estudar um modo de selecionar a quantidade ideal de históricos de usuários a ser utilizada em recomendações, pois assim teremos maiores valores para *precision* e *recall*. Assumimos heuristicamente que os históricos de usuários podem ser divididos em grupos, nossa hipótese neste trabalho é que estes grupos indicam quais os melhores históricos de consultas que devem ser considerados para recomendação de consultas para um usuário. Poderemos assim, também colaborar com a solução de problemas vistos nas demais técnicas apresentadas na Seção 3.2, como a necessidade de uma fase de obtenção de preferências do usuário, conhecimento do esquema do banco de dados e problemas com constantes mudanças nos dados.

4. Desenvolvimento necessário para conclusão

Estamos estudando a relevância dos grupos de usuários para a escolha das consultas que devem ser recomendadas para um usuário que está utilizando o sistema. Na Figura 2, apresentamos como será o funcionamento da técnica proposta em nossa pesquisa. Na etapa 1, diversos usuários acessam o banco de dados que armazena o histórico de consultas realizadas por cada usuário. Na etapa 2, após a extração dos atributos presentes no histórico de consultas dos usuários, é realizado o agrupamento dos usuários utilizando os atributos extraídos e uma técnica de agrupamento. Por fim, são utilizados os atributos extraídos do histórico de consultas e dos grupos de usuários para realizar recomendações de novas consultas na etapa 3.

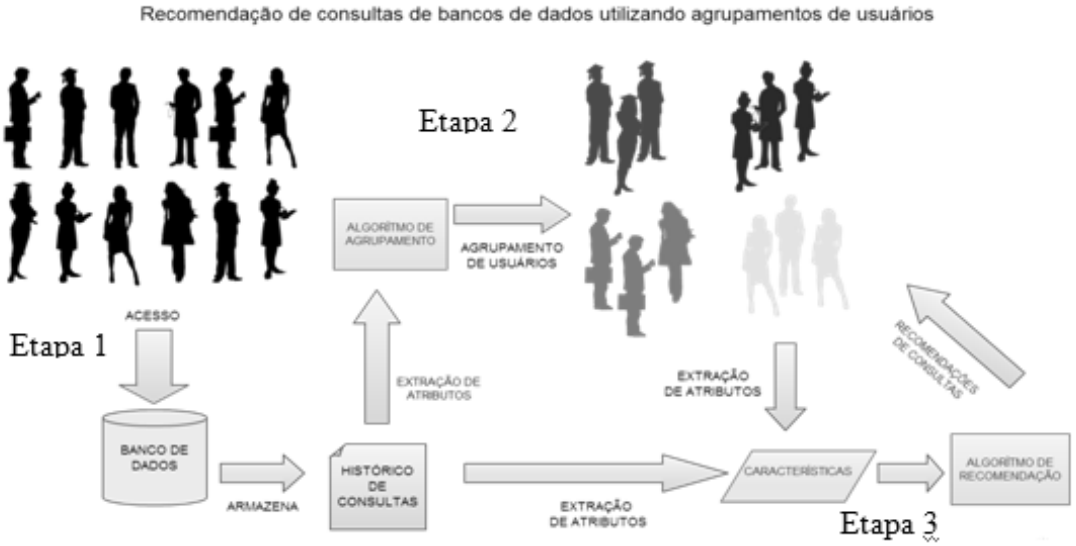

Figura 2. Visão geral da técnica proposta.

Iremos comparar o impacto da criação de grupos nos valores das métricas *precision* e *recall* de técnicas já existentes. Além disso, verificaremos se uma nova técnica de recomendação de consultas que será desenvolvida em nossa pesquisa

apresenta melhoras na qualidade das recomendações de consultas realizadas em banco de dados.

5. Avaliação dos resultados

Assim como descrito na Seção 3.3, repetiremos os experimentos buscando avaliar as métricas observadas após adicionar a etapa de agrupamento de usuários nas técnicas selecionadas e na proposta em nossa pesquisa. Em seguida, serão realizados testes estatísticos buscando demonstrar uma nova técnica para recomendar consultas em banco de dados que retorne recomendações com melhor *recall* e *precision* do que as demais técnicas encontradas no estado da arte.

6. Referências

Koutrika, G. and Ioannidis, Y. (2004). Personalization of Queries in Database systems. In Proceedings of 20th Intl. Conf. On Data Engineering (ICDE). Boston, MA, USA. p. 597-608.

Akbarnejad, J., Chatzopoulou, G., Eirinaki, M., Koshy, S., Mittal, S., On, D., Polyzotis, N. and Varman, J. S. V. (2010). SQL QueRIE recommendations. In Proceedings of the VLDB Endowment. v.3 n.1-2.

Limam, L., Coquil, D., Kosch, H. and Brunie, L. (2010). Extracting user interests from search query logs: A clustering approach. In Proceedings of the 2010 Workshops on Database and Expert Systems Applications (DEXA '10). p. 5-9.

Aligon, J., Golfarelli, M., Marcel, P., Rizzi, S. and Turricchia, E. (2013) Similarity measures for olap sessions. To appear in International Journal of Knowledge and Information Systems (KAIS).

Marcel, P. and Negre, E. (2011). A survey of query recommendation techniques for dataware-house exploration. In Proceedings of the 7th Conference on Data Warehousing and On-Line Analysis (Entrepts de Donnes et Analyse) (EDA'11), p. 119-134.

Khoussainova, N., Kwon, Y., Balazinska, M., Suciu, D. (2010). SnipSuggest: context-aware autocompletion for SQL. In Proceedings of the VLDB Endowment. v.4 n.1, p. 22-33.

Revista The Economist, The Data Deluge, edição do dia 25 de Fevereiro de 2010.

Site do Projeto Genome disponível em http://genome.ucsc.edu/ - acesso em 31/05/2013

Site do Projeto Skyserver disponível em http://cas.sdss.org/ - acesso em 31/05/2013

Stefanidis, K., Drosou, M., Pitoura, E. (2009). You May Also Like results in relational databases. In PersDB, p. 37-42

Yang, X., Procopiuc, C. M. and Srivastava, D. (2009). Recommending join queries via query log analysis. In 25th International Conference on Data Engineering (ICDE 2009), p. 964–975.

Zhang, Z., Nasraoui, O. (2006). Mining search engine query logs for query recommendation, Proceedings of the 15th international conference on World Wide Web, Edinburgh, Scotland, p. 1039-1040.