Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Recomendação de Consultas de Banco de Dados utilizando Agrupamento de Usuários

Márcio de Carvalho Saraiva

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação

Linha de Pesquisa: Sistemas de Informação e Banco de Dados

Carlos Eduardo Santos Pires e Leandro Balby Marinho

(Orientadores)

Campina Grande, Paraíba, Brasil

S243r Saraiva, Márcio de Carvalho.
Recomendação de consultas de banco de dados utilizando agrupamento de usuários / Márcio de Carvalho Saraiva. - Campina Grande, 2014.
67 f.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2014.
"Orientação : Prof. Dr. Carlos Eduardo Santos Pires, Prof. Dr. Leandro Balby Marinho".
Referências.

1. Banco de Dados. 2. Recomendação de Consultas. 3. Agrupamento de Usuários. 4. Perfis de Comportamento. 5. Métricas de Avaliação. 6. Dissertação - Ciência da Computação. I. Pires, Carlos Eduardo Santos. II. Marinho, Leandro Balby. III. Universidade Federal de Campina Grande - Campina Grande (PB). IV. Título

CDU 004.65(043)

"RECOMENDAÇÃO DE CONSULTAS DE BANCO DE DADOS UTILIZANDO AGRUPAMENTOS DE USUÁRIOS"

MÁRCIO DE CARVALHO SARAIVA

DISSERTAÇÃO APROVADA EM 27/08/2014

CARLOS EDUARDO SANTOS PIRES, Dr., UFCG
Orientador(a)

LEANDRO BALBY MARINHO, Dr., UFCG
Orientador(a)

ULRICH SCHIEL, Dr., UFCG
Examinador(a)

ANDREI DE ARAUJO FORMIGA, D.Sc., UFPB
Examinador(a)

CAMPINA GRANDE - PB

Resumo

Os sistemas de banco de dados estão se tornando cada vez mais populares na comunidade científica para suporte à exploração de dados científicos. Neste cenário, os usuários podem não ter o conhecimento necessário sobre o domínio do banco de dados ou não saber formular consultas SQL para análise dos dados. Para resolver este problema surgiram diversos estudos sobre técnicas para recomendação de consultas. Os métodos de recomendação de consultas em banco de dados têm dado ênfase em maximizar apenas a acurácia das recomendações, mas outros aspectos como novidade e diversidade podem ser importantes para recomendações. Nesse contexto, esta pesquisa teve como objetivo melhorar as recomendações de consultas SQL com relação às métricas relevância, novidade, diversidade, quantidade de tabelas novas, *precision* e *recall*. Esse objetivo foi alcançado por meio de uma abordagem para recomendação de consultas utilizando agrupamentos de usuários de banco de dados. Os resultados dos experimentos utilizando históricos de consultas reais do projeto SkyServer mostram que por intermédio da abordagem proposta é possível gerar recomendações de consultas adequadas para cada usuário do banco de dados utilizado. Além disso, foi avaliada a abordagem proposta comparando com técnicas descritas em trabalhos relacionados. As análises realizadas mostram que os valores das métricas estudadas nesta pesquisa são 64,6% maiores na abordagem proposta do que nas técnicas comparadas. Esses resultados possivelmente proporcionarão melhores condições para estudos e trabalhos futuros utilizando agrupamentos de usuários para realizar recomendações de consultas de banco de dados. A abordagem proposta também possibilitou o delineamento de comportamentos de usuários de banco de dados, essa informação colabora para melhor compreensão da interação dos usuários com sistemas de gerenciamento de banco de dados.

Palavras-chave: recomendação de consultas, banco de dados, agrupamento de usuários, perfis de comportamento, métricas de avaliação.

Abstract

Database systems are becoming increasingly popular in the scientific community to support the exploration of scientific data. In this scenario, users may not have the necessary knowledge about the domain of the database or not knowing formulate SQL queries for data analysis. To solve this problem has been emerged many studies about queries recommendation techniques. The recommendation methods of query in database has been emphasis in maximize the accuracy of the recommendations, but other aspects such as novelty and diversity may be important for recommendations. In this context, this research aimed to improve the recommendations of SQL queries regarding the metrics: relevance, novelty, diversity, an amount of new tables, precision and recall. This goal was achieved through an approach to the recommendation of queries using clusters of database users. The results of experiments using real historical queries of the SkyServer project shows that through the proposed approach we can generate recommendations for appropriate queries for each user of the database used. Furthermore, the proposed approach was evaluate comparing with techniques described in related work. The analysis shows that the values of the metrics studied in this research are 64,6% higher in the proposed approach than the techniques compared. These results potentially provide better conditions for studies and future work using groups of users to generate recommendations for database queries. The proposed approach also enabled the design of user behavior with database, this information contributes to better understanding the interaction of users with management systems database.

Keywords: recommendation of queries, database, group of users, behavior profiling, valuation metrics.

Agradecimentos

Agradeço primeiramente a Deus, Pai, Filho e Espírito Santo, e a Nossa Senhora que me encheram de entusiasmo nos momentos que eu mais precisei.

Aos meus pais, Francisco Márcio Eugênio Vieira Saraiva e Solange Braga de Carvalho Saraiva, por todo amor e suporte que me deram para que eu alcançasse meus objetivos na vida e que não me faltaram em momento algum durante a realização deste trabalho, muitas vezes ultrapassando suas limitações para me ajudar.

Aos meus irmãos, Eugênio de Carvalho Saraiva e Fernanda de Carvalho Saraiva Nascimento, por serem meus melhores amigos e mesmo em alguns momentos em que nós não estávamos juntos, nunca me deixaram sentir sozinho, e sempre acreditaram no meu trabalho.

A todos os demais familiares e amigos, pois juntos conseguimos ver que tudo é possível, vibramos com as conquistas de todos, mas também compartilhamos os momentos difíceis e dividimos os fardos da vida.

À minha noiva, Bruna Marques de Almeida, que está sempre ao meu lado, apoiando minhas escolhas, e fazendo tudo que está ao seu alcance para me ajudar a ser um melhor profissional. Por seu amor e carinho que revigoram minhas forças e me auxiliam a seguir meus caminhos.

Aos meus orientadores, Professor Carlos Eduardo Santos Pires e Leandro Balby Marinho, que além de serem expressivos colaboradores para a realização desta pesquisa, me deram significativas lições para a vida.

Aos membros da banca da defesa de dissertação desta pesquisa de mestrado, Professor Ulrich Schiel e Andrei de Araújo Formiga, que com suas observações contribuíram com a melhoria da qualidade deste trabalho.

Aos colegas do Laboratório de Sistemas e Informação (LSI) e do Programa de Educação Tutorial (PET) da Unidade Acadêmica de Sistemas e Computação, onde

este projeto foi desenvolvido, por todas as discussões sobre o trabalho e o apoio na minha vida acadêmica.

À Capes pela concessão da bolsa para pesquisa.

Ao projeto SkyServer pela disponibilização dos dados utilizados nesta pesquisa.

E por fim, à Universidade Federal de Campina Grande pelo ambiente e pelos recursos disponibilizados para meus estudos.

Dedicatória

Aos meus pais que sempre acreditam no meu potencial e me incentivam o contínuo aperfeiçoamento.

Sumário

Glossário

Sigla		Significado
ADC	-	Average Distance to other Candidates - Distância Média para outros Candidatos
BD	-	Banco de dados
IDF	-	Inverse Document Frequency – Inverso da Frequência de Documentos
IFF	-	Inverse Feature Frequency – Inverso da Frequência de Características
IP	-	Internet Protocol - Protocolo de internet
KNN	-	*k*-Nearest Neighbors algorithm – *k*-Vizinhos mais próximos
SGBD	-	Sistema de Gerenciamento de Banco de Dados
SSE	-	Soma dos Erros Quadrados
SQL	-	Structured Query Language - Linguagem de Consulta Estruturada

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Os bancos de dados estão se tornando cada vez mais presentes na vida de usuários e desenvolvedores de diversos sistemas que manipulam os dados armazenados para a descoberta de conhecimento.

Atualmente, os bancos de dados também estão populares na Web e diversas empresas passaram a disponibilizar seus dados para estudos e pesquisas. Por exemplo, o site do projeto UCSC Genome Browser (Genome, 2014), desenvolvido na Universidade de Santa Cruz na Califórnia (UCSC), fornece acesso a um banco de dados sobre genética, o site StackExchange Data Explorer apresenta um banco de dados de respostas para perguntas de diversos temas, enquanto que o site do projeto SkyServer (Skyserver, 2014) disponibiliza um banco de dados sobre astronomia. Na grande maioria dos casos, estes bancos de dados disponibilizados são do tipo relacional.

O acesso a essas bases de dados ocorre por meio de aplicações que possuem uma interface Web e permitem que uma grande quantidade de usuários espalhados pelo mundo envie consultas, normalmente escritas na Linguagem de Consulta Estruturada (SQL). Além disso, os bancos de dados apresentam esquemas complexos contendo uma grande quantidade de elementos (e.g. tabelas, colunas e relacionamentos).

1.1 Motivação

Por mais de 40 anos, os Sistemas de Gerenciamento de Banco de Dados (SGBDs) foram desenvolvidos para prover diversas operações sobre dados. Ao mesmo tempo, ferramentas para manuseio de consultas sobre os dados não tiveram a mesma evolução, uma razão para isso é que as consultas são normalmente emitidas por meio de aplicativos.

Comumente, consultas são depuradas e reutilizadas repetidamente. No entanto, este modo de utilização de consultas está mudando, em diversos trabalhos de pesquisa surge a necessidade de se executar vários tipos de consultas dependendo do resultado de experimentos.

É observado também que muitos cientistas estão armazenando e compartilhando cada vez mais grandes volumes de dados em bancos de dados na Web. Dentre os diversos usuários que acessam bancos de dados na Web, encontramos aqueles que não estão familiarizados com os elementos do esquema do banco de dados e que, por consequência, possuem dificuldade em formular consultas que poderiam recuperar dados relevantes para suas necessidades.

Também é possível observar que novos usuários podem apresentar dificuldades para executar consultas em banco de dados. Em tal caso, tutoriais ou instruções passo-a-passo poderiam reduzir muito o custo nesse processo de aprendizagem. No entanto, manter em dia este tipo de documentação é um processo custoso e demorado.

Estes fatos instigam o estudo de técnicas para recomendação de consultas de banco de dados, que auxiliariam usuários a formular consultas SQL para análise dos dados. Estudar técnicas para recomendação em banco de dados, especialmente para recomendação de consultas, foi visto na literatura como um tópico relevante para estudos (Khoussainova *et al.*,2009; Chatzopoulou *et al.*, 2009), o que contribui, sobremaneira, para a realização deste trabalho de mestrado.

1.2 Limitações na área

Diversos trabalhos têm procurado identificar características em consultas de bancos de dados realizadas por usuários e a exploração dessas características para a recomendação de consultas. Julien Aligon, *et al.* (2014) e Marcel e Negre (2011) revisaram a literatura sobre técnicas de recomendação de consultas e observaram que faltam informações sobre os usuários e seus papéis com relação às consultas realizadas.

Também foi observado que os métodos de recomendação de consultas em banco de dados presentes na literatura têm dado ênfase em maximizar apenas a acurácia das recomendações, mas outros aspectos como novidade e diversidade podem ser importantes para recomendações.

Neste trabalho, verificamos que no domínio estudado não há conhecimento sobre técnicas para agrupamento de usuários de banco de dados em perfis de comportamento de usuários considerando características extraídas do histórico de consultas realizadas pelos mesmos. Este conhecimento pode auxiliar na seleção das consultas a serem recomendadas, uma vez que consultas de usuários com determinados perfis de comportamento podem ser priorizadas.

O conhecimento sobre agrupamentos de usuários também pode melhorar os resultados de métricas que avaliam a qualidade de recomendações como *precision* e *recall*, bastante utilizadas na literatura, além de métricas adaptadas neste trabalho que avaliam a diversidade, novidade, relevância e quantidade de tabelas novas presentes em recomendações de consultas SQL.

1.3 Objetivo do Trabalho

O objetivo geral deste trabalho é melhorar a qualidade das recomendações de consultas SQL por meio de uma abordagem proposta baseada em agrupamentos de usuários construídos segundo atributos extraídos do histórico de consultas realizadas pelos mesmos. Com isso, será possível encontrar informações sobre o comportamento de usuários de banco de dados e melhorar as interações desses usuários com o sistema de banco de dados por meio de recomendações de consultas que podem auxiliar a obtenção de informações relevantes. Para tanto, alguns objetivos específicos foram contemplados, a citar:

1- Caracterizar o problema tratado e identificação de contribuições e trabalhos relacionados;

2- Investigar uma etapa de agrupamento de usuários com a possibilidade de adaptação dessa etapa para técnicas de recomendação de consultas existentes;

3- Definir um algoritmo de recomendação de consultas que utilize as informações de agrupamento de usuários; e

4- Analisar a técnica proposta em relação as suas contrapartidas presentes na literatura.

Este trabalho tem como público-alvo usuários que acessam bancos de dados na Web, que possuem dificuldade em formular consultas que poderiam recuperar dados relevantes para suas pesquisas.

1.4 Relevância

Dado que a revisão do estado da arte, até então realizada, apontou que nenhum dos trabalhos encontrados na literatura busca estudar o agrupamento dos usuários de banco de dados de acordo com perfis de comportamento, por meio de atributos presentes no histórico de consultas dos usuários, pode-se observar a originalidade do estudo presente em nosso trabalho. Tal estudo viabiliza a construção de novas abordagens em sistemas de recomendação de consultas a banco de dados.

Também não é observada em nenhum trabalho a análise da contribuição do comportamento dos usuários para a utilização de um sistema de banco de dados. Esta análise pode auxiliar na tomada de decisão de políticas que beneficiam determinado perfil de comportamento de usuários, assumindo que este perfil realiza mais consultas no sistema e é responsável pela maior parcela do funcionamento do sistema ao retornar dados.

Ao utilizar o perfil de comportamento dos usuários em recomendações de consultas de banco de dados, pode-se obter uma maior precisão nas recomendações,

uma vez que as recomendações podem ser pontuadas levando-se em consideração o perfil do usuário que receberá a recomendação, distinguindo cada usuário dos demais.

O trabalho realizado colabora para as áreas de Sistemas de Recomendação e de Banco de Dados, uma vez que as contribuições serão válidas tanto para técnicas de recomendação de informação, buscando aprimorar a qualidade de recomendações de consultas de banco de dados, como para pesquisas em banco de dados, apresentando comportamentos de usuários que utilizam bancos de dados.

O conhecimento gerado durante os estudos para criação da abordagem proposta poderá ser utilizado em vários cenários. A utilização de grupos para recomendação poderá ser incluída como uma fase para preparar os dados de outras técnicas de recomendação para melhorar métricas de qualidade de recomendação semelhantes às utilizadas nessa pesquisa.

Em cenários de pesquisas que analisam usuários, o método empregado para análise do comportamento de usuários de banco de dados deste trabalho, por intermédio da comparação de atributos e agrupamentos, também poderá ser reutilizado.

1.5 Hipóteses

O conhecimento gerado nesta pesquisa permitiu o desenvolvimento de uma abordagem de recomendação de consultas de banco de dados que torna mais proveitosas as atividades de usuários, familiarizados ou não com o esquema de um banco de dados, que possuam interação com o banco de dados, retornando dados que são importantes para suas consultas.

Para alcançar esse resultado, neste trabalho de dissertação foram criadas as seguintes hipóteses para nortear os estudos realizados:

H1. É possível criar um método para identificar comportamentos de usuários considerando as consultas realizadas pelos mesmos em um banco de dados?

H2. É possível identificar diferenças entre comportamentos de usuários?

H3. Existem métricas para avaliar a qualidade de recomendações de consultas de banco de dados, que podem ser calculadas levando em consideração agrupamentos de usuários?

H4. Existe um modo para recomendar consultas de banco de dados melhorando as métricas: *precision*, *recall,* novidade, diversidade, relevância e quantidade de tabelas novas?

H5. A abordagem proposta apresenta valores maiores para as métricas de avaliação de qualidade estudadas.

1.6 Estrutura da Dissertação

Este trabalho está organizado da seguinte forma:

O Capítulo 2 apresenta a fundamentação teórica sobre Sistemas de Recomendação, Sistemas de Banco de Dados e conceitos sobre técnicas de agrupamento, que serão importantes para melhor compreensão do que estamos propondo.

Os trabalhos relacionados à pesquisa desta dissertação encontram-se no Capítulo 3. Este capítulo contempla explicações de como são realizadas as recomendações de consulta de banco de dados por diversas pesquisas e qual a lacuna de pesquisa deixada por esses trabalhos que foi preenchida por nossos estudos.

O Capítulo 4 contempla a especificação da técnica de recomendação de consultas proposta, os passos necessários para realizar recomendações, como são gerados os agrupamentos de usuários por meio de atributos presentes no histórico de consulta dos mesmos e o modo como são calculadas as métricas utilizadas para

avaliação das recomendações.

O Capítulo 5 mostra os testes realizados para avaliação das métricas citadas no capítulo anterior, e as análises sobre os resultados obtidos após os testes.

Por fim, o Capítulo 6 apresenta as considerações finais. A Seção 6.1 apresenta conclusões e considerações sobre as contribuições propostas, discutindo a relação das mesmas com outros trabalhos da literatura, apontando vantagens e limitações. A Seção 6.2 contempla sugestões de trabalhos futuros identificados ao longo do desenvolvimento desta dissertação.

Capítulo 2

Fundamentação Teórica

Neste capítulo, são apresentados os conceitos relevantes para o entendimento da pesquisa desenvolvida. Primeiramente, são descritos Sistemas de Recomendação, citando suas principais técnicas. Em seguida, é discorrido sobre Sistemas de Gerenciamento de Banco de Dados Relacionais, enfatizando a formulação de consultas. Além disso, são apresentados detalhes sobre parte do ambiente em que este trabalho de mestrado atua, Banco de Dados na Web. Por fim, são descritos conceitos sobre *Clustering* e como o algoritmo K-means, utilizado nesta pesquisa, realiza o agrupamento de objetos.

2.1 Sistemas de Recomendação

Com o aumento da quantidade de informações na Internet, as pessoas passaram a contar com uma grande variedade de opções para obtenção de dados. Comumente, diversos indivíduos possuem pouca experiência para realizar todo tipo de escolhas dentre as diversas alternativas que lhes são apresentadas. Para reduzir as dúvidas e necessidades destes indivíduos frente à escolha entre alternativas, geralmente são utilizadas recomendações que são fornecidas por outras pessoas, seja de forma direta (Maes and Shardanand, 1995) ou através de textos de recomendação, como textos de opiniões de revisores de livros e jornais.

Surgiram então os Sistemas de Recomendação, uma subárea de Recuperação e Filtragem de Informação, que procuram auxiliar as pessoas a encontrar conteúdos de interesse, com base em seu histórico de preferências. Estes sistemas combinam diversas técnicas computacionais para escolher itens personalizados com base no contexto no qual estão inseridos e nos interesses dos usuários.

Com a evolução destes sistemas e o fato deles trabalharem com grandes bases de

informações, diversos estudos realizados na área permitiram que recomendações não triviais fossem realizadas, algumas vezes proporcionando melhores resultados que uma recomendação feita por humanos (Resnick, P. e Varian, H.R., 1997).

Os primeiros trabalhos em sistemas de recomendação podem ser encontrados nas pesquisas em teoria de aproximação, ciências cognitivas, recuperação da informação, teoria de previsões, ciências econômicas e administração (Armstrong, J. S., 2001; e Murthi, B. P. S. e Sarkar, S., 2003). A área de sistemas de recomendação emergiu como uma área de pesquisa independente, nos anos 90.

2.1.1 Técnicas de Recomendação

Existem diversas técnicas para fazer recomendações, algumas comparam as preferências de um usuário com um grupo de outros usuários, outras procuram itens com características parecidas aos que o usuário já demonstrou interesse no passado. É possível coletar de forma implícita ou explícita as preferências do usuário. Na forma implícita, informações são obtidas por meio de interações passadas dos usuários com um sistema ou até mesmo de dados externos, como sua localidade geográfica. A forma explícita utiliza feedbacks efetivos como, por exemplo, notas dadas a um item. As técnicas podem ser classificadas em duas categorias, a partir de como a recomendação é feita:

Baseada em Conteúdo

Um sistema de recomendação baseado em conteúdo recomenda ao usuário itens que sejam semelhantes aos itens que ele preferiu no passado. Esses sistemas são mais simples para dados textuais do que para dados não textuais, e não necessitam de uma grande quantidade de informações sobre o usuário para recomendar itens.

Filtragem Colaborativa

Consiste na recomendação de itens para os usuários que pessoas com histórico

semelhante tiveram acesso no passado. Na Filtragem Colaborativa, os usuários avaliam o quanto gostam de determinados itens, e analisando essas avaliações, os sistemas determinam qual será a avaliação de um usuário para um item ainda não avaliado.

A vantagem desse tipo de recomendação é a diminuição da incidência de recomendações repetitivas e apresentam resultados positivos (LINDEN, G. *et al,* 2003). A desvantagem é a necessidade de muitas avaliações dos usuários sobre itens do sistema para funcionar precisamente.

Também foram criados os sistemas híbridos, que combinam as duas abordagens mencionadas, tentando explorar suas vantagens e superar as desvantagens.

Neste trabalho de dissertação, apresentamos uma abordagem apoiada na técnica de filtragem colaborativa, pois as consultas recomendadas são provenientes de usuários que possuem comportamento semelhante ao usuário que receberá as recomendações.

A seguir na Seção 2.2, descrevemos os sistemas nos quais as recomendações geradas por nossa abordagem serão utilizadas.

2.2 Sistemas de Gerenciamento de Banco de Dados

Bancos de dados são coleções de dados que foram organizadas para se ter mais eficiência durante pesquisas e assim gerar mais informações e conhecimento sobre os dados (Elmasri, Navathe, 2011). Os bancos de dados podem armazenar informações sobre pessoas, produtos, ou qualquer outra coisa. São de grande importância para instituições de pesquisa, como universidades, laboratórios e empresas.

Por mais de duas décadas, e ainda hoje, se tornaram peça fundamental para o desenvolvimento de sistemas de informação. Normalmente existem por vários anos sem alterações em sua estrutura.

Os Sistemas de Gerenciamento de Bancos de Dados (SGBD) são ferramentas muitas vezesutilizadas para manipulação de dados e informações para aplicações nas quais temos vários usuários e a organização entre estes é necessária.

Atualmente, os Sistemas de Gerenciamento de Bancos de Dados se tornaram parte do cotidiano da sociedade. Diariamente, diversas atividades envolvem alguma interação com os bancos de dados como, por exemplo, em um banco para se efetuar um depósito ou retirar dinheiro, ou ao realizar compras ou consultas de informações de produtos pela internet.

Nos últimos anos, os avanços tecnológicos geraram inúmeras novas aplicações para os Sistemas de Gerenciamento de Banco de Dados, como o gerenciamento de imagens, vídeos e informações geográficas.

2.2.1 Tabelas

Os bancos de dados podem conter várias tabelas. Uma tabela de banco de dados é semelhante a uma planilha, uma vez que os dados são armazenados em colunas e linhas. A principal diferença entre armazenar os dados em planilhas e armazená-los em um banco de dados é a maneira como os dados são organizados.

Cada linha em uma tabela é chamada de registro ou tupla, que apresenta locais onde os itens individuais de informações são armazenados. As tuplas possuem um ou mais campos, que correspondem às colunas da tabela. Cada campo deve pertencer a um determinado tipo de dados, como texto, número ou data.

2.2.2 Consultas

Consultas em banco de dados são operações que podem executar várias funções diferentes, porém a função mais utilizada é recuperar dados específicos de tabelas, visualizadas em tuplas. Geralmente, os dados que os usuários de banco de dados desejam ver estão distribuídos em várias tabelas, e às vezes necessitam de consultas avançadas para exibi-los. Frequentemente as consultas servem como fonte de registros para formulários e relatórios para pesquisas científicas e operações empresariais.

Existem diversas formas e linguagens para se realizar uma consulta, mas a linguagem de consulta estruturada ou SQL é a linguagem de pesquisa declarativa padrão

para banco de dados relacional.

Diferente de outras linguagens de consulta a banco de dados que especificam o caminho para alcançar resultados (linguagens procedurais), uma consulta SQL especifica a forma do resultado (linguagem declarativa).

As recomendações realizadas pelo trabalho desta dissertação apresentam consultas SQL uma vez que a base de dados utilizada possui apenas consultas deste tipo e a técnica de recomendação proposta faz uso de trechos de código presentes neste tipo de consultas.

Em consultas SQL, esses trechos de código são chamados de cláusulas. As cláusulas são condições utilizadas para definir os dados que deseja selecionar em uma consulta. As cláusulas básicas em SQL e que são consideradas nessa pesquisa são:

- **SELECT**: Obrigatória em uma consulta SQL, é utilizada para listar os dados das colunas que serão projetados na consulta;
- **FROM**: É obrigatória e juntamente com a cláusula *SELECT* formam a base de consultas SQL. Nesta Cláusula é especificada a fonte das informações que se vai selecionar os registros dos dados, podendo ser apenas uma ou várias; e
- **WHERE**: Não é obrigatória. Esta cláusula especifica as condições que restringem os dados obtidos por meio de operações que verificam se cada registro satisfaz ou não as condições especificadas;

Exemplo de uma consulta SQL:

```
SELECT nome FROM Aluno WHERE idade>18
```

Nesta consulta são retornados os nomes presentes na coluna "nome" da tabela "Aluno" nos quais os valores presentes na coluna "idade" sejam maiores que 18.

A seguir, na seção 2.3, são apresentados conceitos a respeito do ambiente no qual os históricos de consulta de usuários utilizados neste trabalho de dissertação executaram consultas SQL.

2.3 Banco de Dados na Web

A web representa, nos dias de hoje, um repositório universal de dados, onde a quantidade de sites existentes e o volume disponível de dados é muito grande. A informação na Web muitas vezes está distribuída e desorganizada, o que faz com que os sites fiquem cada vez mais complexos para se navegar, dificultando a manipulação e a consulta de dados de múltiplas fontes.

Nesse contexto, a recuperação de informação na Web consiste, basicamente, em busca por palavras-chave e navegação. Com o tempo, diversos sites e aplicações da web vêm necessitando de ferramentas para facilitar a gerência de dados. A Web oferece inúmeras oportunidades para uso da tecnologia de bancos de dados como, por exemplo, o fato de bancos de dados serem geralmente bem projetados, os dados seguem uma estrutura rígida e são manipulados em um ambiente controlado.

Analisando o contexto da Web, é possível verificar o surgimento de questões não abordadas tradicionalmente na área de Banco de Dados. A dinamicidade da web não permite validar situações estáveis, nas quais é assumido uma quantidade fixa de fontes de dados a serem integradas.

Tradicionalmente em um banco de dados que não está na web, a quantidade de dados quando não é fixa, é considerada de alteração pouco frequente. Porém pesquisas que utilizam banco de dados na Web normalmente consideram que esta alteração pode ocorrer diversas vezes. (LIMA, F., *et.al*, 1999)

Também é possível observar em bancos de dados na Web que a quantidade de usuários que fazem acesso à SGBDs utilizados para gerenciar esses bancos de dados não é fixa e pode variar com o tempo.

Na seção a seguir são apresentados conceitos sobre *Clustering,* importantes para compreensão do modo que são realizados os agrupamentos de usuários presentes nesta pesquisa de dissertação.

2.4. Clustering

Devido ao número elevado de informações, classificar ou agrupar dados em categorias tem se tornado uma atividade comum (Backer, 1995). Agrupamentos possibilitam meios para identificar indivíduos com comportamentos isolados em um sistema e o impacto de ações de grupos, avaliar a dimensionalidade, e sugerir hipóteses referentes ao inter-relacionamento de entidades que pertencem ou não ao mesmo grupo. Esta tarefa é possível graças a um mecanismo chamado clusterização, clustering ou análise de agrupamento.

Clustering é uma técnica de aprendizado não-supervisionado (Jain, Dubes, 1988) dentro da área de Mineração de Dados, que permite encontrar automaticamente agrupamentos de dados segundo as semelhanças entre eles.

Na Figura 1 é apresentado um exemplo de clustering. Na parte (a), um conjunto de dados de entrada, onde cada elemento é representado pelo símbolo 'x'. Na parte (b) é apresentado o resultado de um algoritmo de clustering realizado sobre esse conjunto de dados de entrada, com cada elemento 'x' sendo rotulado com um número identificador do grupo (cluster) a que pertence ao final do clustering.

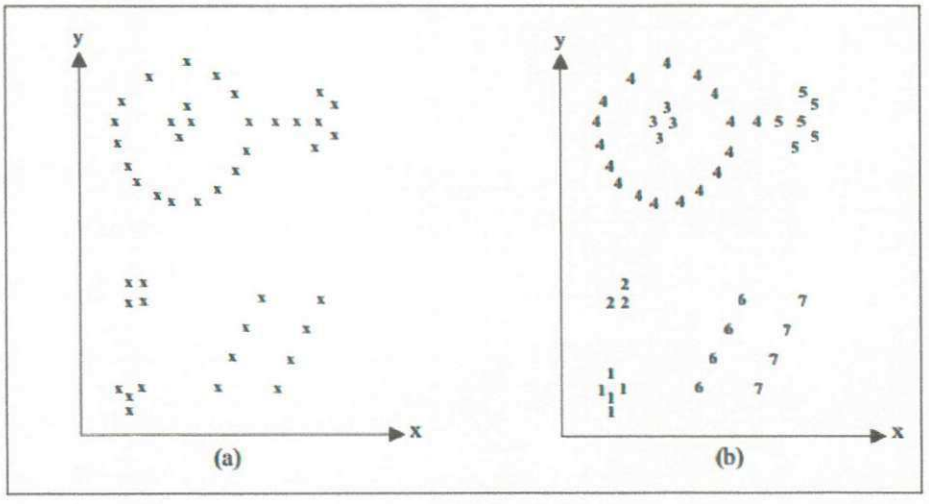

Figura 1 – Exemplo de Clustering: Os dados de entradas são apresentados na parte (a) e os 7 clusters construídos são apresentados na parte (b). (Jain et al., 1999).

Na maioria dos sistemas que criam agrupamentos (*clusters*) o usuário deve escolher primeiramente o número de grupos a serem detectados, porém existem algoritmos que conseguem detectar de forma automática o número de grupos presentes em uma base de dados, há também outros que necessitam apenas o número mínimo de clusters.

Dois passos importantes para algoritmos de clusterização são as definições de medidas de similaridade entre dois agrupamentos e do cálculo da distância entre os elementos dos agrupamentos. Existem vários algoritmos que fazem agrupamento, eles são classificados como hierárquicos ou particionais (também conhecidos como sequenciais ou iterativos).

Neste trabalho, o algoritmo de clusterização K-means (MacQueen, J. B., 1967) foi escolhido para geração dos agrupamentos de usuários, por ser um dos algoritmos de clusterização mais utilizados na literatura, descrito na próxima seção.

2.4.1 Algoritmo K-means

K-means é um algoritmo de *Clustering* que busca particionar n objetos em k clusters, nos quais os objetos que pertencem ao mesmo cluster são mais semelhantes entre si e menos semelhantes a objetos em outros clusters.

O algoritmo fornece uma classificação de informações de acordo com os próprios dados e automaticamente apresenta uma classificação automática sem a necessidade de nenhuma supervisão humana.

Para classificar os objetos e gerar grupos, o algoritmo executa uma comparação entre cada valor numérico referente aos dados por meio de uma distância, neste trabalho foi utilizada a distância euclidiana, descrita na Equação 1.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik}, x_{jk})^2}$$

(1)

para k = 1, ..., n quantidade de componentes dos padrões

Onde,

d = distância entre os padrões xi e xj no espaço de dimensão

x_{ik} = componente k do padrão x_i

x_{jk} = componente k do padrão x_j

Após o cálculo das distâncias o algoritmo encontra os centróides para cada um dos grupos. O algoritmo padrão do K-means utiliza uma técnica de refinamento interativo, isto é, conforme o algoritmo vai iterando, o valor de cada centróide é refinado pela média dos valores de cada atributo de cada ocorrência que pertence a este centroide. Quando o valor do centróide não é alterado o algoritmo para as iterações e os agrupamentos são observados.

O algoritmo é composto pelos seguintes passos:

1. Colocar K pontos no espaço representado pelos objetos que estão sendo agrupados. Estes pontos representam centros dos Grupos iniciais.
2. Atribuir a cada objeto para o grupo que tem o centróide mais próximo.
3. Quando todos os objetos foram atribuídos, recalcular as posições dos centróides K.
4. Repetir as etapas 2 e 3 até que os centróides não se movam.

2.5. Considerações Finais

Após a apresentação dos conceitos presentes neste capítulo é possível ter uma melhor compreensão do ambiente em que este trabalho está inserido e de quais conhecimentos e ferramentas são necessárias para o desenvolvimento da abordagem para recomendação

de consultas proposta desta dissertação.

No capítulo seguinte é introduzida uma revisão de literatura que evidencia a relevância da pesquisa realizada neste trabalho de mestrado, bem como apresenta fatos pertinentes para demostrar as contribuições atingidas pelo estudo realizado.

Capítulo 3

Revisão de Literatura

Este capítulo apresenta pesquisas sobre técnicas de recomendação de consultas que abordam diversos aspectos, fazendo uma análise comparativa com o estudo desenvolvido neste trabalho de dissertação.

Foi realizado um levantamento do estado da arte, referente às pesquisas que são relacionadas a 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. Por meio de nossos estudos, podemos afirmar que há apenas as tentativas de Chatzopoulou *et al.* (2011) e Stefanidis *et al.*, (2009) de formalizar conceitos de 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 e 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, enquanto Stefanidis *et al.* utiliza usuários x consultas, Chatzpoulou *et al.* faz uso de sessões de acesso x tuplas recuperadas.

Utilizando também as tuplas recuperadas para realizar a reformulação de consultas, encontramos os trabalhos de Tran Q.T. e Chan C.Y (2010) e Sarkas, N. et al (2009), nos quais consultas são "relaxadas" (são retiradas condições para retornar mais dados) ou "restringidas" (são acrescentadas condições para diminuir os dados retornados) dependendo da quantidade de resultados utilizando a reescrita de termos ou a expansão de consultas para melhorar as métricas *recall* e o *precision* da consulta

original.

Por fim, é possível encontrar as pesquisas de Sarma, A.D *et al.* (2010) e Tran Q.T. e Chan C.Y (2010) que constroem consultas que recuperam tuplas equivalentes às recuperadas pela consulta original, analisando o retorno da consulta realizado pelo usuário no momento de sua execução.

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, a matriz sessões de acesso dos usuários x 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 por meio de uma interface gráfica os fragmentos mais prováveis de *Q'* a partir de um fragmento inicial *Q*, a técnica completa automaticamente consultas para formular pesquisas nos dados.

Buscando identificar características dos usuários de banco de dados, o trabalho de Lyes Limam *et al.* (2010) extrai o interesse (partes mais acessadas 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 *et al.* (2009), 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 apresentam padrões de comportamento de usuários, para formular perfis de comportamento, por meio de características extraídas do histórico de consultas realizadas pelos mesmos.

Padrões de comportamento podem 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.

3.1 Comparando Técnicas de Recomendação de Consultas de Banco de Dados

Primeiramente, as técnicas estudadas foram separadas de acordo com o tipo de entrada utilizado, apresentadas na Tabela 1. Utilizamos esse critério, pois uma das contribuições da abordagem sugerida neste trabalho é a utilização de agrupamentos de usuários como entrada para algoritmos de recomendação de consultas.

Tabela 1. Técnicas de recomendação de consultas estudadas de acordo com o tipo de entrada utilizado

<table>
<tr><th>Usuários x Consultas</th><th>Usuários x Tuplas Recuperadas</th><th>Usuários x Fragmentos de Consultas</th><th>Usuários x Esquema e Dados</th><th>Usuários x registros de preferências do usuário</th></tr>
<tr><td rowspan="4">Stefanidis et al. (2009)</td><td>Chatzopoulou et al. (2011)</td><td rowspan="3">Akbarnejad et al. (2010)</td><td rowspan="3">Lyes Limam et al. (2010)</td><td rowspan="3">Yang et al.(2009)</td></tr>
<tr><td>Tran Q.T. e Chan C.Y (2010)</td></tr>
<tr><td>Sarma, A.D et al. (2010)</td></tr>
<tr><td>Sarkas, N. et al (2009)</td><td>Khoussainova et al. (2010)</td><td>Zhang e Nasraoui (2006)</td><td>Koutrika e Ioannidis (2004)</td></tr>
</table>

Os trabalhos que utilizam usuários x registros de preferências do usuário necessitam de uma fase de obtenção de preferências, 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], essa fase de obtenção de preferências se torna inadequada, pois são necessários vários registros de preferências, que não são ordinariamente obtidos dos usuários.

Por esse motivo, as técnicas apresentadas nos trabalhos de Yang *et al.* (2009), e de Koutrika e Ioannidis (2004) não foram selecionadas para a elaboração desta pesquisa de mestrado.

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

[1] Por exemplo, o esquema do banco de dados utilizado nesta pesquisa, disponibilizado pelo projeto *SkyServer*, contém 91 tabelas.

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 selecionados três tipos de técnicas de recomendação de consultas, usuários x consultas, usuários x tuplas recuperadas e usuários x matriz de fragmentos de consultas, que utilizam como parâmetro de entrada o histórico de consultas dos usuários. Desses tipos de técnica de recomendação, foram selecionadas as técnicas citadas nos trabalhos mais recentes para serem utilizadas em nossos estudos. São elas: Akbarnejad et. al. (2010), Stefanidis et. al. (2009) e Chatzopoulou et. al. (2011).

A seguir, é apresentada a descrição das técnicas selecionadas para o estudo realizado neste trabalho de dissertação.

3.2 Técnicas Selecionadas

Foram selecionadas três técnicas de recomendação de consultas que utilizam como parâmetro de entrada o histórico de consultas dos usuários e que podem utilizar a base de dados disponibilizada para este estudo (histórico de consultas do projeto SkyServer).

3.2.1 Akbarnejad *et al.* (2010)

A técnica tem dois mecanismos de recomendação distintos, cada um usando um conceito diferente de similaridade. O primeiro mecanismo de recomendação, define a similaridade entre dois usuários em termos de suas necessidades de informação.

Esse mecanismo de recomendação é baseado em tuplas e identifica quais dados do banco de dados foram acessados por consultas do usuário atual e recupera os usuários que têm explorado os mesmos dados no passado. Estes, juntamente com as consultas do usuário atual, definem um conjunto de dados do banco de dados cobertos até o momento por consultas do usuário.

O conjunto final de recomendações consiste em consultas que tem melhor

cobertura da base de dados. No entanto, duas consultas podem ser semanticamente semelhantes, mas recuperar dados diferentes devido a alguma filtragem de condições. Logo, os autores propuseram um segundo mecanismo de recomendação baseado em fragmentos.

Neste mecanismo as consultas realizadas pelos usuários são divididas em fragmentos (atributos, tabelas, junções e predicados), assim é criada uma matriz usuário x fragmento. A partir de um fragmento de consulta, o sistema procura, com base na frequência de uso em consultas, outros fragmentos do histórico de usuários que são similares entre si que compõem um conjunto dos fragmentos mais importantes para um usuário, aqueles que mais são utilizados em consultas do usuário.

O trabalho de Akbarnejad *et al.* (2010) retorna recomendações utilizando as consultas armazenadas que contém fragmentos do conjunto de fragmentos criado.

Algoritmos propostos

Os autores assumem que as consultas de cada usuário visualizam um subconjunto de dados do banco de dados que é relevante para a análise do que o usuário deseja executar. Este subconjunto é modelado como um resumo da sessão S_i para o usuário i. Foi utilizado {1,. . . , H} para designar o conjunto de usuários que acessaram o banco de dados no passado com base no qual as recomendações são geradas e 0 para identificar o usuário atual. Para gerar recomendações, a técnica proposta estende o resumo S_0 do usuário ativo para um resumo "previsto" S_0^{pred}. Este resumo captura o grau previsto de interesse do usuário atual com respeito a todas as partes da base de dados, incluindo aqueles que o usuário ainda não explorou e serve como uma entrada para a geração de recomendações.

A estrutura da técnica proposta é composta por três componentes: (a) a construção de um resumo da sessão de S_i para cada usuário i, (b) o cálculo de um

resumo "previsto" S_0^{pred} para o usuário ativo, baseado na resumos dos últimos dos usuários ativos do usuário, e (c) a geração de consultas com base no S_0^{pred}. Essas consultas serão apresentadas ao usuário como recomendações. A seguir os detalhes de cada passo dos dois mecanismos utilizados para recomendação.

Mecanismo baseado em tuplas

- **Resumos de sessão**: O resumo de sessão Si é definido como um vetor de pesos de tuplas que abrange todas as tuplas do banco de dados. O peso de cada elemento do vetor representa a importância da respectiva tupla na exploração realizada pelo usuário i. Usando os resumos das sessões de usuários passados, é possível definir a matriz de sessão-tupla, como no caso da matriz usuário-item de sistemas de recomendação na Web, será utilizada como entrada em um processo de filtragem colaborativa.
- **Calculando o resumo de sessão "previsto"**: Da mesma forma que os resumos de sessão, o resumo estendido S_0^{pred} é um vetor de pesos de tupla. Para calcular este resumo, é assumido a existência de uma função *sim*(Si, Sj), que mede a similaridade entre dois resumos e assume valores no intervalo [0, 1]. Usando esta função, é calculado o resumo estendido como uma soma ponderada dos resumos existentes:

$$S_0^{pred} = \sum_{0 \le i \le H} (sim(S_0, S_i) x S_i) \quad (2)$$

A função de similaridade *sim* pode ser realizada com qualquer métrica baseada em vetor, como a medida de similaridade do cosseno.

- **Gerar recomendações**. O passo final é gerar consultas que cobrem as tuplas interessantes em S_0^{pred}. A fim de proporcionar aos usuários recomendações

intuitivas, de fácil compreensão, são utilizadas consultas de usuários passados. É atribuído para cada consulta Q passada a importância no que diz respeito ao S_0^{pred}, calculado como *rank* (Q, S_0^{pred}.) = *sim* (S_Q, S_0^{pred}). Assim, uma consulta possui alto valor se cobre tuplas importantes de S_0^{pred}. Consultas no topo do ranking são, então, retornadas como a recomendação.

Mecanismo baseado em fragmentos

- **Resumos de sessão**: Este mecanismo baseia-se na semelhança de pares de fragmentos de consulta (atributos, tabelas, junções e predicados). É necessário identificar os fragmentos que co-aparecem em várias consultas representada por diferentes usuários. O vetor de resumo da sessão S_i para um usuário *i* é constituída por todos os fragmentos de consulta θ de consultas passadas do usuário. Considerando Q_i como a representação do conjunto de consultas feitas por *i* e *F* a representação do conjunto de fragmentos de consulta distintas registradas nos históricos de consulta. Para um dado fragmento $\theta \in F$, a sua importância em sessão S_i é representada por $S_i[\theta]$. Os autores definem $S_Q[\theta]$ como um peso variável ou binário que representa a importância de θ em consulta Q de uma sessão. Então, S_i é definido como uma soma($S_i = \sum_{Q \in Q_i} S_Q$) ou *Or-ed*($V_{Q \in Q_i} S_Q$).

- **Calculando o resumo de sessão "previsto"**: Usando os resumos das sessões dos últimos usuários e um vetor semelhança, foi construído o (| F | × | F |), matriz fragmento-fragmento, que contém todas as semelhanças *sim* (*p*, θ), $p, \theta \in F$. O resumo de sessão "previsto", modelado por S_0^{pred}, representa a importância estimada de cada fragmento de consulta no que diz respeito ao

comportamento S_0 do usuário ativo. De forma semelhante a abordagem item-a-item de filtragem colaborativa de sistemas de recomendação web, os autores empregaram as similaridades de fragmento-a-fragmento que são calculadas na etapa anterior, como ilustrado na Equação 3:

$$S_0^{pred}[\theta] = \frac{\sum_{p \in R} S_0[p] * sim(p,\theta)}{\sum_{p \in R} sim(p,\theta)} \quad (3)$$

onde R representa o conjunto top-k de fragmentos similares de consultas ($k \leq |F|$).

- **Gerar recomendações**: Uma vez que o resumo "previsto" S_0^{pred} tenha sido computado, os fragmentos que tenham recebido o maior peso são selecionados (top-n fragmentos de Fn). Então todas as consultas passadas Q, $Q \in U_i Q_i$ recebem uma classificação QR com base em uma métrica normalizada que mede o número de fragmentos de consulta comuns de cada consulta Q à lista de top-n. Finalmente, as consultas classificados top-m são usados como a recomendação final.

A seguir, será apresentado um exemplo utilizado pelos autores para ilustrar o funcionamento da técnica proposta em seu trabalho. O trabalho interage com o banco de dados do projeto SkyServer e utiliza históricos de usuários do passado.

Dado um usuário que precisa formular consultas SQL complexas, mas não tem a experiência necessária. Suponha que o usuário precisa realizar alguma análise específica, que requer o uso de agregações, a fim de retornar os resultados desejáveis. O usuário começará a submeter consultas mais simples. Como por exemplo as consultas:

```
SELECT field FROM PHOTOOBJ

SELECT ra FROM PHOTOOBJ
```

Se, no entanto, um outro usuário no passado passou por este processo, e as respectivas sessões de usuário são registradas nos logs de consulta, o sistema será capaz recomendar a consulta:

```
SELECT ra, field FROM PHOTOOBJ WHERE ra > 200 and field < 100
```

para o usuário atual, uma vez que essa consulta possui fragmentos que aparecem em grande quantidade nas consultas realizadas pelo usuário atual.

3.2.2 Stefanidis *et al.* (2009)

Neste trabalho, os autores apresentam três abordagens fundamentalmente diferentes para computar resultados da ferramenta proposta em sua pesquisa. A primeira, denominada *corrente-estado*, utiliza os resultados da consulta atual e o conteúdo da base de dados. A segunda, denominada *baseada em histórico*, utiliza o histórico de consultas de usuários para sugerir tuplas que são resultados de tanto consultas passadas similares ou resultados de consultas representada por usuários similares. O último, chamado de *fontes externas*, usa informações de recursos externos ao banco de dados, como a web.

Porém, os autores se concentraram na abordagem do *corrente-estado* e apresentaram um novo método para calcular os resultados de sua ferramenta, utilizando análise local e global.

Análise Local

Durante a análise local de um resultado de consulta *R(q)*, busca-se descobrir padrões para recomendar resultados. No trabalho de Stefanis *et al.* (2009), os padrões são vistos com valores de atributos que aparecem frequentemente. Para quantificar as aparições de valores de atributo em *R(q)*, foi definida a matriz valor-frequências *MR(q)*.

Há uma linha de *MR(q)* para cada atributo A1. . . , Am de *R(q)* e uma coluna para cada valor do atributo distinto V1,. . . , Vn contabilizando a quantidade de aparições em consultas. *MR(q)* (i, j) contém o número de ocorrências de Vj por Ai em *R(q)*.

A seguir, um exemplo presente nos trabalhos dos autores ilustrando uma parte da matriz valor-frequências para um usuário interessado em filmes com o ator Lee Phelps.

Tabela 2. Matriz de valor-frequencias para um usuário interessado em filmes com o ator Lee Phelps.

	Policial	Detetive	Tira	Garçom	Guarda
Papel	36	24	23	22	13

Dada uma consulta *q* e os correspondentes valores de frequências na matriz *MR(q)*, o objetivo do trabalho é apresentar ao usuário um conjunto de resultados com cardinalidade *p*. Tais resultados são computados com relação aos valores de atributos mais frequentes em *R(q)* como especificados pelo *MR(q)*. Em particular, são localizados os k elementos de *MR(q)* com os valores mais altos e para cada um desses elementos são construídas consultas apropriadas para recuperar resultados interessantes. Para maior clareza na notação, os autores consideram a matriz *M'R(q)* para *M'R(q)* (i, j) = *MR(q)* (i, j) se *MR(q)* (i, j) pertence ao k, k> 0, os valores dos atributos mais frequentes e *M'R(q)* (i, j) = 0 se não. Cada elemento contribui com um número de resultados segundo a função F, descrita na Equação 4.

$$F(i,j) = \frac{M'_{R(q)}(i,j)}{\sum_i \sum_j M'_{R(q)}(i,j)} \cdot p \qquad (4)$$

Para o exemplo anterior utilizando p=10 e k=2, seriam recomendados 6 filmes em que o ator Lee Phelps participou como "Policial" e quatro filmes como "Detetive".

Análise global

Na análise global busca-se utilizar propriedades de banco de dados durante a recomendação de resultados. Neste trabalho, os autores consideram estatísticas específicas mantidas no banco de dados utilizado para suas pesquisas. Por esse motivo, não foi utilizado esse método para comparação das métricas entre a abordagem descrita neste trabalho de dissertação e o trabalho de Stefanidis *et al.* (2009). Nesta pesquisa de mestrado foi utilizada apenas a análise local.

Os autores 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.

Após criada essa matriz, o algoritmo proposto consegue estipular a utilidade de cada consulta realizada por outros usuários no banco de dados comparando a similaridade da mesma com a consulta mais recente do usuário que receberá a recomendação. Essa similaridade é calculada utilizando a matriz valor-frequência, onde consultas com termos com frequência mais próxima são mais similares.

Com a utilidade calculada, são retornadas como recomendações as consultas que mais se assemelham com as consultas com maior utilidade.

3.2.3 Chatzopoulou *et al.* (2011)

O trabalho de Chatzopoulou *et al.* (2011) é semelhante ao trabalho de Akbarnejad *et al.* (2010), porém apresenta algumas divergências no cálculo do resumo de sessão "previsto", S_0^{pred}.

A técnica apresentada pelos autores gera recomendações para S_0 em duas etapas. Em primeiro lugar, é computado o S_0^{pred}, que capta a importância das diferentes características de consulta para S_0 usuário. Em S_0^{pred} podem conter tuplas que já aparecem nas consultas de S_0, mas também novos que o usuário ainda não utilizou. O

S_0^{pred} é utilizado no segundo passo do fluxo de trabalho criado para recomendação, como "semente" para gerar recomendações.

Diferente do trabalho de Akbarnejad *et al.* (2010), o resumo previsto é calculado como S_0^{pred} = f (α, S0, S1, ..., Sn), onde f é uma função que combina a informação dos resumos S0, S1,. . . , Sn. O Parâmetro α é um fator de mistura, o qual controla a importância de S0 (a sessão do usuário atual) em relação a S1,. . . , Sn (as sessões de outros usuários).

Os autores também utilizam dois mecanismos para gerar as recomendações, um baseado em tuplas e outro baseado em fragmentos. No mecanismo baseado em tuplas, o calculo do S^{pred} é feito da seguinte forma: $S^{pred} = \alpha \cdot S_0 + (1 - \alpha) \cdot \sum_{i=1..n} sim(S_i, S_0).S_i$, onde $sim(S_i, S_0)$ é uma métrica de similaridade entre os dois vectores (por exemplo, similaridade do cosseno) e α é o fator para se misturar a sessão do usuário atual com sessões de usuários passados.

Tendo calculado S^{pred}, as recomendações são consultas de usuários que utilizaram o banco de dados no passado que são mais similares com as consultas presentes em S^{pred}, encontradas por meio da função $sim(S_Q, S^{pred})$, onde S_Q é um resumo de sessão que contém apenas a consulta Q.

O segundo mecanismo, baseado em fragmentos, funciona da mesma forma descrita por Akbarnejad *et al.* (2010). Por esse motivo, neste trabalho de dissertação foi utilizado somente o mecanismo baseado em tuplas da técnica proposta por Chatzopoulou *et al.* (2011) para realização de estudos, uma vez que o mecanismo baseado em fragmentos está presente no trabalho de outros autores também selecionados para estudo.

Nessa pesquisa é utilizada como entrada sessões de acesso x tuplas que foram retornadas pelas consultas realizadas pelos usuários.

Porém, para utilizar nesta pesquisa de mestrado apenas técnicas que podem usar a base de dados disponibilizada, foi desenvolvida a técnica proposta pelos autores

substituindo as tuplas que seriam usadas no algoritmo por consultas dos usuários. Dessa maneira, temos o trabalho de Chatzopoulou *et al.* (2011) com a entrada sessão de acesso x consulta.

3.3 Considerações Finais

Por serem trabalhos recentes, podemos concluir a existência de necessidade pesquisas neste campo e possíveis trabalhos futuros para contribuir com os resultados encontrados pelos pesquisadores de cada um desses trabalhos.

Todos os trabalhos utilizam como entradas para as técnicas de recomendação os históricos de consultas dos usuários, porém nenhum dos trabalhos tenta unificar as informações dos históricos de consultas dos usuários por meio de agrupamentos de usuários ou realiza algum ensaio com outros tipos de agrupamentos.

Também não é observada nenhuma tentativa de extração de características das consultas dos usuários com o objetivo de analisar o comportamento dos mesmos, definindo padrões de interação no sistema de banco de dados.

Por fim, podemos reparar que nos trabalhos estudados faltam informações sobre métricas para avaliação da novidade e da diversidade das consultas recomendadas.

Na Figura 2 é apresentado um resumo do modo que são realizadas recomendações de consulta em sistemas encontrados na literatura.

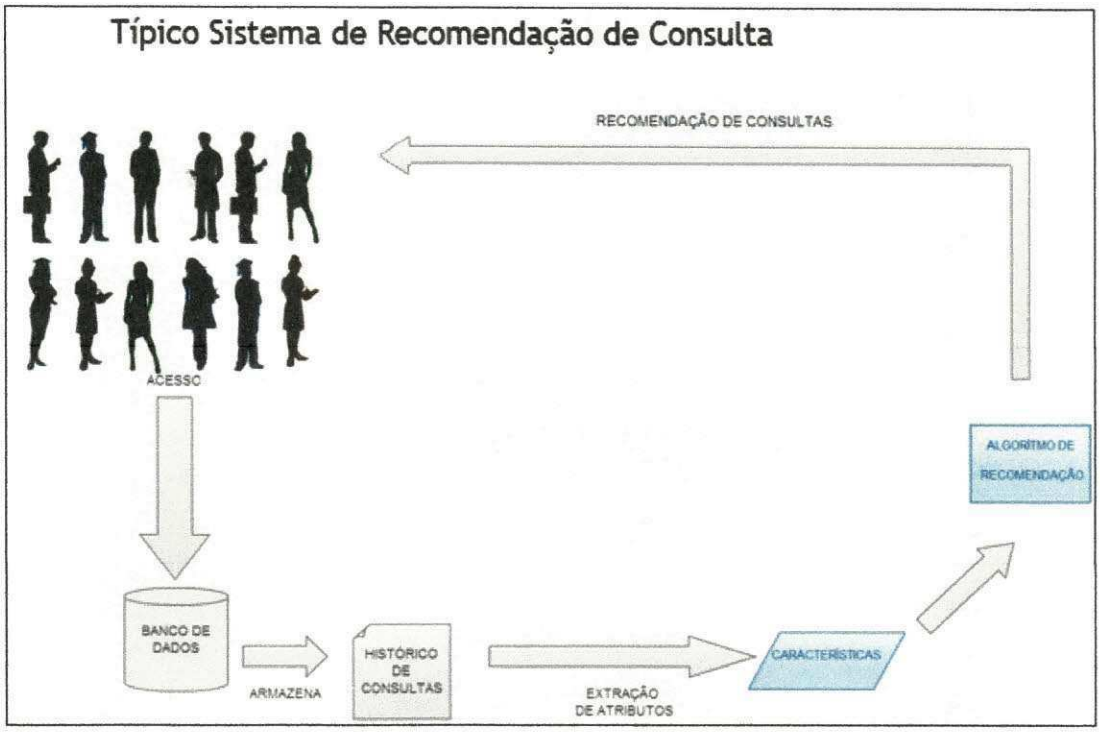

Figura 2 - Típico Sistema de Recomendação de Consultas

Por intermédio desses sistemas, vários usuários acessam o banco de dados e submetem consultas. O banco de dados, por sua vez, retorna informações e armazena o histórico de consultas realizadas por cada usuário.

Posteriormente, os sistemas de recomendação de consultas extraem atributos sobre os acessos dos usuários. Nas técnicas selecionadas esses atributos podem ser as consultas realizadas, tuplas retornadas ou fragmentos de consultas. Esses atributos representam características do usuário que são utilizadas para realizar recomendações de consultas.

No Capítulo 4 é apresentada a abordagem proposta por esta pesquisa de mestrado para recomendação de consultas, visando contribuir com as lacunas observadas nos trabalhos relacionados que foram selecionados para estudo. No Capítulo 5 as técnicas selecionadas juntamente com a abordagem proposta são avaliadas utilizando métricas descritas na seção 4.2.

Capítulo 4

Abordagem proposta para Recomendação de Consultas

Neste capítulo é apresentada a abordagem proposta para recomendações de consultas SQL para usuários de banco de dados. A abordagem é baseada em agrupamentos de usuários construídos segundo características extraídas do histórico de consultas realizadas pelos mesmos.

Na seção 4.1 é mostrado, em duas etapas, como são geradas as recomendações de consultas por meio da abordagem proposta. Em seguida, nas subseções 4.1.2 e 4.1.3, são detalhadas cada uma das etapas, especificando como são tratadas as entradas de cada etapa e os artefatos disponíveis após cada etapa.

Finalmente, na seção 4.2 são apresentadas as métricas relevância, quantidade de tabelas novas, novidade, diversidade, *precision e recall*, as quais são utilizadas para avaliação das recomendações geradas pela abordagem proposta.

4.1 Passos da abordagem proposta

Em nossa pesquisa, estudamos o impacto da utilização de grupos de usuários na escolha das consultas que devem ser recomendadas para um usuário que está utilizando um sistema de banco de dados.

Na Figura 3, é apresentado uma visão geral da abordagem proposta. Primeiramente, diversos usuários acessam o banco de dados e submetem consultas. O banco de dados, por sua vez, retorna os resultados e armazena o histórico de consultas realizadas por cada usuário.

Após a extração de informações sobre o acesso de usuários por meio de consultas (nesta pesquisa foram utilizados os nomes das tabelas acessadas pelos usuários) presentes no histórico de consultas dos usuários, são gerados grupos de usuários com base nas informações extraídas através de uma técnica de agrupamento.

Por fim, são utilizados os atributos dos históricos do usuário e informações dos grupos de usuários da etapa 1 para realizar recomendações de consultas na etapa 2. A

seguir, são apresentados dois exemplos de histórico de consultas que serão utilizados para detalhar as etapas 1, 2, nas subseções 4.1.1 e 4.1.2.

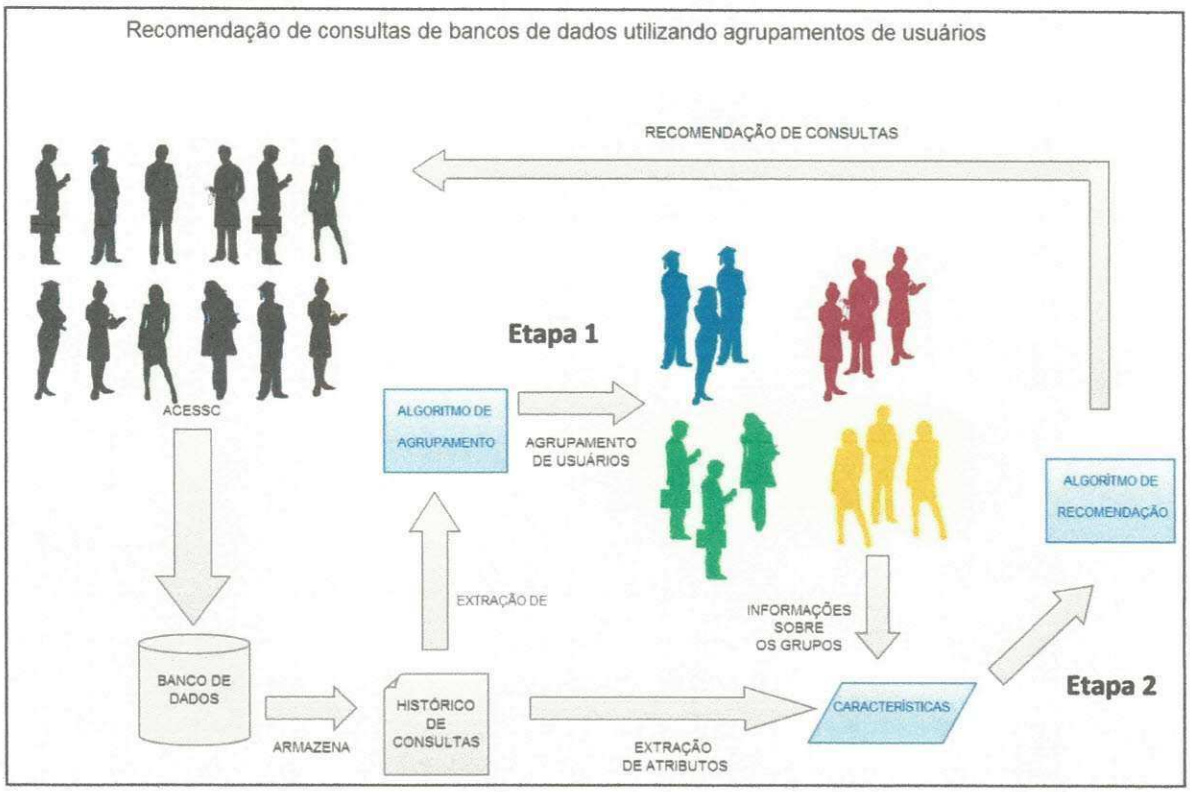

Figura 3 - Visão Geral da Abordagem Proposta.

4.1.1 Armazenamento do Histórico de Consultas dos Usuários

As consultas de cada usuário são armazenadas em arquivos. Após o armazenamento, são obtidos os históricos de consulta de cada usuário. Assim, é possível extrair informações do texto das consultas escritas pelos mesmos, como a quantidade de condições utilizadas e a quantidade de tabelas acessadas em uma consulta. Também são calculados os atributos que foram elencados neste trabalho para caracterizar um padrão de acesso do usuário, como: média de consultas realizadas por hora utilizando o sistema de banco de dados, quantidade de horas que o usuário utilizou o sistema e quais tabelas foram utilizadas nas consultas.

Esses atributos foram escolhidos para serem utilizados na pesquisa desta dissertação por serem simples de obter a partir de elementos presentes no histórico de consulta dos usuários: data de submissão da consulta, endereço IP (Internet Protocol ou Protocolo de internet) da máquina que a submeteu, e texto da consulta escrito na

linguagem SQL.

A seguir, são apresentados dois exemplos de históricos de consultas que retornam dados de imagens espaciais do banco de dados disponibilizado pelo projeto SkyServer, definidos como HC1 e HC2, de usuários denominados U1 e U2, respectivamente. Esses usuários e históricos são utilizadas ao longo do capítulo para facilitar a explicação das etapas necessárias para recomendação de consultas de acordo com a técnica proposta em nossa pesquisa.

O histórico de consulta HC1 contém as seguintes consultas:

a) SELECT * FROM PHOTOOBJ WHERE ra > 234

b) SELECT ra FROM PHOTOOBJ

c) SELECT ra, field FROM PHOTOOBJALL WHERE ra > 200 and field < 100

d) SELECT n.objID, p.objid FROM FGETNEARBYOBJEQ as n, PHOTOPRIMARY as p', WHERE n.objID=p.objID

O histórico de consulta HC2 contém as seguintes consultas:

a) SELECT p.objid p.run, p.rerun, p.camcol, p.field FROM PHOTOPRIMARY as p

b) SELECT s.mjd,s.plate,s.fiberid from PHOTOOBJ AS p JOIN BESTDR8..Specobj AS s ON s.bestobjid = p.objid where s.z >= 0.010 and s.z< 0.015 and p.fracdev_g > 0.95 and p.fracdev_r > 0.95 and p.fracdev_i > 0.95 and p.r < 16.5 and s.class 'GALAXY'

c) SELECT s.mjd,s.plate,s.fiberid from Photoobj AS p JOIN BESTDR7..Specobj AS s ON s.bestobjid = p.objid where s.z >= 0.050 and s.z< 0.055 and p.fracdev_g > 0.95 and p.fracdev_r > 0.95 and p.fracdev_i > 0.95 and p.r < 16.5 and s.SpecClass=dbo.fspecClass('GALAXY')

d) SELECT s.mjd,s.plate,s.fiberid from PHOTOOBJ AS p JOIN BESTDR7..Specobj AS s ON s.bestobjid = p.objid where s.z >= 0.045 and s.z< 0.050 and p.fracdev_g > 0.95 and p.fracdev_r > 0.95 and p.fracdev_i > 0.95 and p.r < 16.5 and s.SpecClass=dbo.fspecClass('GALAXY')

Utilizando esses históricos de consultas, são calculadas cada um dos atributos citados, que serão armazenados para serem utilizados pela técnica de recomendação proposta e na avaliação dos resultados.

• ***A1 - Número médio de consultas realizadas por hora*:** razão entre a quantidade de consultas presentes no histórico do usuário e a diferença em horas entre o primeiro acesso e o último acesso do usuário no sistema.

Exemplo: se um usuário realizou uma consulta em 11/10/2011 às 11h00, outra às 22h00 e outra em 12/10/2011 às 15h00, conclui-se que três consultas foram realizadas em um período de 28 horas. Logo, o número médio de consultas realizadas por hora será 3/28 = 0,1 consulta por hora.

• ***A2 - Quantidade de horas em que o usuário utilizou o sistema de banco de dados:*** consiste no intervalo de tempo em horas entre a primeira e a última consulta realizada pelo usuário no sistema de banco de dados.

Exemplo: com base no exemplo do atributo anterior, o usuário gastou 28 horas utilizando o sistema. Este atributo também pode ser chamado de quantidade de horas desde que o usuário teve o primeiro acesso ao sistema de banco de dados.

• ***A3 - Número médio de tabelas utilizadas por consulta:*** razão entre a quantidade de tabelas presentes na cláusula "FROM" das consultas SQL realizadas pelo usuário e a quantidade de consultas presentes no histórico de consultas do usuário.

Exemplo: considere o histórico de consultas HC1. Neste histórico estão presentes quatro consultas, três com uma tabela na cláusula "FROM" e uma com duas tabelas. Logo, o número médio de tabelas utilizadas por consulta será 5/4 = 1,25 tabela por consulta. Deve-se considerar o nome das tabelas com repetição, pois dessa maneira são evitados valores menores que 1 no resultado desse atributo, uma vez que nas consultas presentes nos históricos de consultas dos usuários utilizados nesta pesquisa sempre acessam pelo menos uma tabela.

• ***A4 - Número médio de condições utilizadas por consulta:*** razão entre a quantidade de condições encontradas na cláusula "WHERE" das consultas SQL realizadas pelo usuário e a quantidade de consultas presentes no histórico de consultas do usuário.

Exemplo: utilizando o mesmo histórico adotado no exemplo do atributo anterior, temos a consulta "a" com uma condição (ra>234), a consulta "c" com duas condições (ra>200 e field<100) e a consulta "d" com uma condição (n.objID=p.objID) na cláusula "WHERE". Assim, o número médio de condições utilizadas por consulta para este histórico de consultas é 4/4 = 1 condição por consulta.

- ***A5 - Tabelas acessadas:*** nome das tabelas distintas presentes no histórico de consultas do usuário.

Exemplo: utilizando o histórico adotado para cálculo do atributo *A3*, temos as seguintes tabelas acessadas: PHOTOOBJ, PHOTOOBJALL, FGETNEARBYOBJEQ e PHOTOPRIMARY.

- ***A6 – Quantidade de acessos a uma tabela:*** quantidade de vezes que um usuário executou consultas a uma tabela presente no histórico de consultas do usuário.

Exemplo: novamente utilizando o histórico adotado para cálculo do atributo *A3*, temos as seguintes quantidades de acesso paras as tabelas: PHOTOOBJ foi acessada duas vezes, PHOTOOBJALL, FGETNEARBYOBJEQ e PHOTOPRIMARY foram acessadas uma vez.

Na subseção 4.1.2 é apresentado como são gerados os agrupamentos de usuários utilizando a informação sobre as tabelas acessadas. Dos atributos citados, apenas as tabelas acessadas serão utilizadas como entrada para algoritmos geradores de agrupamentos, pois somente este atributo é observado para cálculo das métricas utilizadas para avaliação, que serão apresentadas na seção 4.2 desta dissertação.

Os atributos que informam sobre quantidade de horas no sistema, média de consultas realizadas por hora, quantidade de condições e tabelas acessadas por consulta serão utilizadas na análise dos resultados, na seção 5.4.

4.1.2 Gerando os Agrupamentos de Usuários

Para gerar os agrupamentos dos usuários de um sistema de banco de dados, extraem-se sem repetição as tabelas que foram acessadas pelo usuário, que estão presentes no histórico de consultas realizadas de cada usuário do sistema.

Em seguida, o conjunto de tabelas utilizadas por cada usuário deve ser substituído por um vetor de tamanho igual ao número de tabelas presentes no banco de dados, onde cada espaço do vetor é preenchido com "0" para indicar que o usuário não acessou a tabela referida ou "1" para indicar que o usuário acessou a tabela. Dessa forma, é criado um vetor de características que representa as tabelas acessadas por cada usuário e é possível calcular a distância entre usuários para gerar os

agrupamentos.

O atributo A6 ("quantidade de vezes que um usuário executou consultas a uma tabela") não foi utilizado, pois este atributo apresenta grande variedade de valores, o que poderia impactar na criação de vários grupos com poucos usuários. Além disso, foi assumido nos estudos desta dissertação que usuários que acessam as mesmas tabelas no esquema do banco de dados possuem o mesmo interesse, independentemente do número de acessos de cada um dos usuários.

Na Tabela 3, temos um exemplo de vetores que representam as tabelas acessadas por cada usuário presente no exemplo utilizado para este capítulo.

Tabela 3 - Amostra dos vetores que indicam as tabelas acessadas por cada usuário.

Usuário	PHOTOOBJ	PHOTOOBJALL	SPECOBJ	PHOTOPRIMARY	FGETNEARBYOBJEQ
U1	1	1	0	1	1
U2	1	0	0	1	0

Por fim, os vetores criados são utilizados como entrada em uma técnica de clusterização para gerar os grupos de usuários com base nas tabelas acessadas. As informações sobre os agrupamentos gerados e em qual grupo cada usuário se encontra são armazenados para serem utilizados posteriormente na técnica de recomendação de consulta especificada na seção seguinte.

4.1.3 Recomendação de consultas utilizando agrupamentos de usuários

O algoritmo de recomendação de consultas proposto tem como entrada o usuário que irá receber as recomendações (usuário alvo) e os agrupamentos gerados no processo descrito na seção anterior.

Na primeira fase do algoritmo, é identificado qual o grupo que o usuário pertence e é selecionado o grupo que irá apresentar melhor valor para a métrica de qualidade selecionada pelo usuário. O algoritmo obtém esta informação por meio das informações geradas no momento do agrupamento dos usuários, conforme descrito na seção anterior deste trabalho de dissertação.

Para avaliar os valores de algumas métricas estudadas, citadas na seção 4.2, é

necessário testar o algoritmo de recomendação de consultas utilizando cada um dos grupos encontrados para realizar recomendações para todos os usuários presentes em nossa base.

Na segunda fase, são selecionados os usuários de um grupo que são mais próximos do usuário alvo, levando em consideração as tabelas que o usuário acessa por meio de suas consultas presentes no histórico de consultas de cada usuário, utilizando os vetores criados na seção 4.1.2 e o cálculo da distância Euclidiana em uma técnica que seleciona os k-usuários mais próximos. Nos testes realizados com grupos diferentes do grupo do usuário alvo, são selecionados aleatoriamente os usuários, pois é assumido que a distância entre usuários de grupos diferentes é infinita. Dessa forma, todos os usuários de outros grupos têm a mesma distância ao usuário alvo.

Finalmente, na terceira fase os históricos de consultas dos usuários escolhidos são reunidos em um novo histórico de consultas e são escolhidas de forma aleatória dez consultas presentes nesse histórico, que serão retornadas como recomendações para o usuário alvo. Foi adotado a quantidade de dez recomendações por ser um valor recorrente na literatura, como por exemplo em "listas top-10", para diminuir o esforço do usuário na leitura das recomendações.

Na Figura 4, é apresentada a representação em pseudocódigo do algoritmo proposto de recomendação de consultas utilizando agrupamento de usuários.

```
ALGORITMO PROPOSTO(usuarioAlvo,agrupamentosGerados,metricaDesejada)

1 agrupamentoRecomendacao <-
  selecionaAgrupamento(usuarioAlvo,agrupamentosGerados,metricaDesejada)

2 k <- 3

3 usuariosMaisProximos <-
  selecionaMaisProximos(k,usuarioAlvo,agrupamentoRecomendacao)

4 historicoUtilizado <-selecionaHistoricoDeConsultas(usuarioMaisProximo)

5 quantidade <- 10

6 recomendacoes <-
  selecionaConsultas(historicoUtilizado, quantidade)

7 retorna recomendacoes
```

Figura 4 - Pseudocódigo do algoritmo proposto de recomendação de consultas utilizando agrupamento de usuários.

O algoritmo de recomendação de consultas recebe como entradas o usuário alvo que irá obter as consultas geradas pelo algoritmo e os agrupamentos gerados por uma técnica de clusterização, conforme citado na seção 4.1.2.

Na linha 1 do pseudocódigo, o algoritmo seleciona qual agrupamento vai ser utilizado para gerar as recomendações. Neste trabalho, o algoritmo de recomendação foi executado diversas vezes e foram selecionados todos os agrupamentos para gerar recomendações para todos os usuários da base de dados utilizada.

Da linha 2 e 3, são escolhidos os usuários mais próximos do usuário alvo (neste trabalho foi definido empiricamente o valor de três usuários mais próximos, "k=3") utilizando os vetores criados na seção 4.1.4 e a distância euclidiana. Na linha 4, o algoritmo cria um histórico de consultas a partir da união dos históricos de consultas dos três usuários selecionados.

Por fim, nas linhas 5 e 6 são selecionadas de forma aleatória dez consultas que formarão uma lista de recomendações que é retornada para o usuário na linha 7 do pseudocódigo.

4.2 Métricas utilizadas para avaliação dos agrupamentos na recomendação de consultas

A pesquisa em sistemas de recomendação tem, historicamente, dado ênfase em maximizar apenas a relevância dos itens recomendados (Vargas e Castells, 2011), isto é, o quanto eles estão de acordo com os interesses do usuário alvo.

Entretanto, relevância apenas pode não ser suficiente para garantir a eficácia e a utilidade das recomendações (McNee *et al.*, 2006; Vargas e Castells, 2011). Por exemplo, considere uma lista de consultas recomendadas a um usuário u em que todos as consultas estão relacionadas ao histórico de consultas de u, porém todos as consultas retornam os mesmos dados.

Além disso, suponha que estas consultas já tenham sido utilizadas por u. Em ambos os casos, embora as consultas recomendadas tenham relevância máxima, elas são menos interessantes e úteis do que uma lista de consultas mais diversificada que traga dados e informações novas para o usuário.

Isso é particularmente importante porque em banco de dados na Web, diversos usuários podem formular consultas diferentes para visualizarem os mesmos dados.

Vargas e Castells (2011) definem novidade como o quão diferente um item é dos itens observados em um dado contexto. Nesse trabalho de dissertação é analisado a novidade de uma consulta no contexto de todas as recomendações feitas para todos os usuários do sistema. Uma consulta é nova se ela não é frequente na aplicação, o que pode ser estimado pelo inverso da popularidade da consulta. Na seção 4.2.2, será apresentado um exemplo para cálculo da novidade de consultas.

Por sua vez, a diversidade de uma lista de consultas recomendadas refere-se ao quão diferente cada item é dos demais (Vargas e Castells, 2011). Neste trabalho de mestrado, a diversidade de uma lista de consultas é estimada pela quantidade de tabelas diferentes acessadas pelas consultas da lista. Será apresentado um exemplo para cálculo da diversidade de consultas na seção 4.2.3.

Assim, consultas mais novas e diversas ajudam a capturar várias características do banco de dados. Portanto, novidade e diversidade são também aspectos importantes para recomendações de consultas, além de sua relevância.

Neste trabalho é proposta a métrica quantidade de tabelas novas, que avalia a quantidade de tabelas presentes nas consultas recomendadas que não foram acessadas por nenhuma consulta do histórico do usuário. Os cálculos de todas as métricas utilizadas para avaliação das recomendações realizadas nessa pesquisa serão descritos e exemplificados nas seções 4.3.1 à 4.3.5.

4.3.1 Relevância

Nesta pesquisa de dissertação é definido que uma consulta recomendada é relevante a um usuário, quando essa faz acesso a uma tabela que está presente em alguma consulta do histórico desse usuário.

Utilizando os históricos de consulta apresentados como exemplo neste capítulo, é possível exemplificar o que são consultas relevantes de acordo com a métrica relevância descrita nesta seção com o seguinte exemplo:

Seja L1 uma lista de consultas recomendadas que possui as consultas C_1 :

```
SELECT p.run,p.field FROM PHOTOOBJ as p WHERE (p.ra > 234.499
AND p.ra < 234.833) AND (p.dec > 4.64708 AND p.dec < 4.98042)
```

e C_2 :

```
SELECT ra From SPECOBJ WHERE ra BETWEEN 194 and 195 AND dec
```

```
BETWEEN 2 and 3
```

C1 é relevante para U1 e C2 não é relevante, pois a consulta C2 não faz acesso a uma tabela presente em alguma consulta do histórico de U1.

Para uma lista de consultas recomendadas, a relevância é calculada utilizando um método similar ao método de Jaccard (MacQueen, 1967), como descrito na Equação 7.

Dada a função *f(x)* que retorna as tabelas acessadas por consultas presentes em uma lista de consultas *x*:

$$f(x) = tabelas\,acessadas$$

$$R\big(f(L), f(Hu)\big) = \frac{|f(L) \cap f(Hu)|}{|f(L)| + |f(Hu)|} \tag{7}$$

Em que *L* representa a lista de consultas recomendadas ao usuário *u* e *Hu* o histórico de consultas de *u*, a interseção calcula a quantidade de consultas de *L* que acessam tabelas presentes no *Hu*. Em seguinda, essa quantidade de consultas é dividida pela quantidade de consultas da lista *L* somadas a quantidade de consultas presentes no histórico do usuário. Para o usuário U1 do exemplo utilizado, a lista L1 tem relevância igual a 1/6, uma vez que apenas uma consulta da lista é relevante para U1, temos duas consultas recomendadas na lista e quatro consultas no histórico de U1.

4.3.2 Novidade

A novidade de uma consulta se refere a frequência que ela ocorre nas recomendações realizadas para os usuários. Quanto menor a frequência de uma consulta nas recomendações para os usuários maior será o valor da novidade, o que pode ser estimado pelo inverso da popularidade da consulta.

Como as consultas realizadas em um sistema de banco de dados acessado por vários usuários, na expressiva quantidade dos casos são diferentes umas das outras, em nossos estudos a novidade de uma consulta é calculada pela novidade da tabela acessada pela consulta.

Novidade está relacionada à probabilidade de que uma tabela não tenha sido observada anteriormente, e portanto quanto menor a popularidade de uma tabela nas recomendações realizadas, mais "nova" ela será. Assim, é utilizado a métrica *Inverse*

Feature Frequency (Belém *et al.*, 2012) para estimar a novidade de uma tabela. O IFF é uma adaptação do tradicional *Inverse Document Frequency* (IDF) que no nosso estudo é considerada a frequência de uma tabela em um histórico de consultas de um usuário específico.

Dado o número N de consultas do histórico de consultas de usuários de cada grupo, o IFF de um termo candidato t é definido na Equação 8:

$$IFF(t) = \log \frac{N+1}{f_t^c + 1} \tag{8}$$

Em que, f_t^c é o número de consultas *c* que contêm a tabela t como parte da cláusula FROM. O valor 1 é somado ao numerador e ao denominador para tratar de novos termos que não aparecem como tabelas nos históricos de consulta dos usuários. A ideia básica da métrica adaptada ao contexto de recomendação de consultas é que tabelas muito frequentes tendem a ser também recomendações muito frequentes, mais óbvias e repetitivas.

A seguir é apresentado o calculado do IFF de uma consulta para o usuário U2 do exemplo utilizado para este capítulo apresentado na seção anterior.

```
SELECT ra, modelMag_r FROM PHOTOPRIMARY WHERE (ra between
54.02095 and 55.16302) and modelMag_r<21 and type=6
```

$$\text{IFF(PHOTOPRIMARY)} = \log\frac{4+1}{1+1}\log\frac{4+1}{1+1} = 0{,}3979.$$

Calculando por meio da Equação 9, a média dos valores de novidade de cada consulta da lista de consultas recomendadas, é possível estimar a novidade da lista de consultas *L* utilizando a função f que retorna as tabelas acessadas em x, dividindo pela quantidade de consultas em *L*:

$$f(x) = tabelas\ acessadas$$

$$MédiaNovidade = \frac{\sum IFF(f(L))}{|L|}$$

(9)

4.3.3 Diversidade

Outro possível aspecto desejável em uma lista de recomendações é a diversidade, isto é, a diversidade de uma lista de consultas recomendadas refere-se ao quão diferente cada consulta é das demais. Estima-se a diversidade de uma lista de consultas como a média da dissimilaridade das tabelas acessadas entre cada par de consultas da lista, de modo que um conjunto de consultas que acessam as mesmas tabelas tenha baixa diversidade.

No contexto de recomendação de consultas, procura-se evitar recomendações redundantes, tais como listas repletas de consultas sinônimas. De forma similar ao que foi feito nos estudos de Sigurbjörnsson e R. van Zwol (2008), neste trabalho de dissertação é estimada a diversidade de uma tabela t em relação a uma lista *R* de outras tabelas de consultas candidatas à recomendação pela média das distâncias entre *t* e cada termo de *R*. Assim, é definida a distância média para outros candidatos (ADC) como na Equação 10:

$$ADC\ (t,R) = \frac{1}{|R|}\sum_{t=1}^{|R|} dist(t, Ri)$$

(10)

Em que $dist(t, Ri)$ é uma medida de dissimilaridade entre as tabelas *t* e R_i. Para estimar essa distância, cada tabela é representada pelo conjunto de consultas em que ela aparece e estimado com a Equação 11 a distância entre duas tabelas t1 e t2 pela diferença relativa entre esses dois conjuntos:

$$dist(t1,t2) = \frac{|C_1 - C_2|}{|C_1 \cup C_2|}$$

(11)

Em que C_1 e C_2 são os conjuntos de consultas que contêm as tabelas t_1 e t_2, respectivamente. Essa métrica corresponde ao complemento do coeficiente de Jaccard entre os conjuntos C_1 e C_2. Nesse trabalho de dissertação é atribuído o valor

$dist(t1, t2)$= 1 (máximo de distância), caso ambos C_1 e C_2 forem vazios.

Exemplo: A tabela PHOTOPRIMARY é utilizada em 200 consultas da base de dados, enquanto que a tabela PHOTOOBJALL aparece em 50 consultas, logo a distância entre as duas tabelas é igual: dist(PHOTOPRIMARY, PHOTOOBJALL) = 150/250 = 0,6

A diversidade de uma lista de consultas recomendadas é calculada como a média da distância entre todas as tabelas presentes nas consultas dessa lista.

4.3.4 Tabelas Novas

Também é calculada a quantidade de tabelas novas (TN) presentes na lista de consultas recomendadas para um usuário, isto é, que não estejam presentes no seu histórico de consultas, de acordo com a Equação 12:

$$TN = |Tr - (Tr \cap Th)| \tag{12}$$

Em que, Tr representa o conjunto de tabelas presentes na lista de consultas recomendadas e Th o conjunto de tabelas presentes no histórico de consultas do usuário. Como exemplo temos o cálculo de TN para o usuário U2 utilizando a seguinte lista de consultas recomendadas:

```
SELECT ra, modelMagErr_z FROM PHOTOPRIMARY WHERE (ra
between 54.02095 and 55.16302) and (dec between 44.42176
and 45.58491) and modelMag_r<21 and type=6

SELECT p.type,p.ra,p.dec,p.u,p.g,p.r,p.i,p.z FROM
PHOTOOBJALL as p WHERE p.ra between 211.520 AND 211.686 AND
p.dec between 61.362 AND 61.528 AND p.r between 18.28 AND
20.28 AND p.type = 6

Select ra From SPECOBJ WHERE ra BETWEEN 194 and 195 AND dec
BETWEEN 2 and 3
```

Tem-se o valor TN = 3 – 1 = 2 para o usuário U2.

Dessa maneira, é possível quantificar em uma métrica o quanto uma lista de consultas auxilia no processo de exploração do esquema do banco de dados utilizados, isto é, quão novas tabelas podem ser acessadas pelo usuário. Quanto maior for o valor

de TN, mais a lista de consultas recomendadas auxilia na exploração do esquema.

4.3.5 Precision e Recall

No contexto de recuperação de informação, *precision* e *recall* são definidos em termos do conjunto de objetos retornados pelas recomendações e o conjunto de objetos relevantes. Neste trabalho são consideradas relevantes as tabelas que podem ser encontradas no histórico de consultas dos usuários, conforme descrito anteriormente.

Precision é a fração das consultas recomendadas que são relevantes para o usuário, calculado de acordo com a Equação 13.

Dada a função *f(x)* que retorna as tabelas acessadas por consultas presentes em uma lista de consultas *x*:

$$f(x) = tabelas\ acessadas$$

$$Precision\big(f(L), f(Hu)\big) = \frac{|f(L) \cap f(Hu)|}{|f(L)|} \tag{13}$$

Em que *L* representa a lista de consultas recomendadas ao usuário *u* e *Hu* o histórico de consultas de *u*, a interseção calcula a quantidade de consultas de *L que* acessam tabelas presentes no *Hu.* Essa quantidade de consultas é dividida pela quantidade de consultas recomendadas pela lista.

Seja L2 uma lista de consultas recomendadas com as consultas

```
SELECT p.run,p.field FROM PHOTOOBJ as p WHERE (p.ra > 234.499
AND p.ra < 234.833) AND (p.dec > 4.64708 AND p.dec < 4.98042)

Select ra From SPECOBJ WHERE ra BETWEEN 194 and 195 AND dec
BETWEEN 2 and 3

SELECT ra, modelMagErr_z FROM PHOTOPRIMARY WHERE (ra between
54.02095 and 55.16302) and (dec between 44.42176 and 45.58491)
and modelMag_r<21 and type=6
```

Para o usuário U1 do exemplo utilizado, a lista L2 tem valor de *precision* igual a 2/3, uma vez que duas consultas da lista são relevantes para U1 e são recomendadas 3 consultas na lista.

Recall em recuperação de informação é a fração dos objetos que são relevantes pelos objetos relevantes que foram retornados pela recomendação. A seguir na Equação 14 é apresentado o cálculo dessa métrica.

Dado a função f(x) que retorna as tabelas acessadas por consultas de x:

$$f(x) = tabelas\ acessadas$$

$$Recall\big(f(L), f(Hu)\big) = \frac{|f(L) \cap f(Hu)|}{|f(Hu)|} \tag{14}$$

No exemplo descrito anteriormente, para o usuário U1 a lista L2 tem *recall* igual a 2/4, uma vez que duas consultas da lista são relevantes para U1 e no histórico de consultas do usuário U1 são encontradas 4 tabelas relevantes.

4.3.6 Aplicando cada métrica estudada

Utilizando cada um dos agrupamentos encontrados por meio do método descrito na seção 4.1.4, são realizadas recomendações para cada um dos usuários presentes em nossa base de dados, que são armazenadas em conjuntos de arquivos texto.

Para cada conjunto de recomendações são calculados os valores das métricas descritas na seção anterior. Dessa maneira, são observados os valores de relevância, quantidade de tabelas novas, novidade e diversidade para recomendações realizadas por cada agrupamento para cada usuário.

Por fim, com o conhecimento de qual grupo cada usuário pertence, é calculado a média dos valores apresentados para as métricas estudadas para cada um dos agrupamentos.

Assim, são obtidas as médias de relevância, quantidade de tabelas novas, novidade e diversidade para cada agrupamento, utilizando como entrada na técnica proposta cada um dos agrupamentos e todos os usuários do sistema de banco de dados com usuários alvo. As análises referentes aos valores encontrados são apresentadas na seção 5.2.

Capítulo 5

Implementação

5.1 Ambiente utilizado para execução dos experimentos

Os experimentos realizados nesta pesquisa foram executados em um notebook com a seguinte configuração:

- Processador Intel Core 2 Duo;
- 4GB Memória RAM; e
- HD 320GB.

Os programas utilizados foram:

- NetBeans IDE 7.2.1
- Weka 3.6
- RStudio, versão 0.97.551

Os dados utilizado nessa pesquisa foram disponibilizados pelo projeto SkyServer, um arquivo no formato CSV, com dados sobre os usuários do banco de dados na web do projeto durante o período de 03/2010 à 07/2011.

Foram utilizados os dados das colunas "yy", "mm", "dd", "hh", "mi", "ss", referentes a data (ano, mês e dia) e o horário (hora, minuto e segundo) em que o usuário executou uma consulta. Também foram utilizados os dados das colunas "clientIP", que representa o endereço IP do computador que o usuário utilizou para acessar o banco de dados do projeto SkyServer, e os dados da coluna "statement", que representam as consultas executadas pelos usuários. Na Figura 4 é apresentada uma amostra dos dados utilizados nesta pesquisa de mestrado.

	A	B	C	D	E	F	G	H
1	yy	mm	dd	hh	mi	ss	clientIP	statement
2	2011	3	8	11	5	22	83.102.250	select name,
3	2011	3	8	11	5	15	220.233.214	SELECT top 1
4	2011	3	8	11	5	13	220.233.214	SELECT top 1
5	2011	3	8	11	5	10	83.102.250	select name,
6	2011	3	8	11	1	48	83.102.250	SELECTp.ra,p.
7	2011	3	8	11	1	40	111.69.149	select name,

Figura 5 - Amostra dos dados utilizados.

5.2 Implementações para gerar agrupamentos de usuários

Utilizando um programa escrito em Java (Java, 2014) no ambiente de desenvolvimento NetBeans IDE 7.2.1 (NetBeans, 2014), capaz de selecionar os nomes das tabelas presentes entre os campos "FROM" e "WHERE" das consultas SQL realizadas pelos usuários, foram gerados os vetores que indicam as tabelas acessadas por cada usuário, conforme processo descrito na seção 4.1.3.

Esses vetores foram usados como entrada do algoritmo de agrupamento K-means provido pela ferramenta Weka (Weka, 2014). Neste algoritmo de agrupamento é necessário definir o valor de k que representa a quantidade clusters, que os objetos serão divididos.

Para definirmos o valor de *k* ideal para a base de dados disponibilizada para esta pesquisa, utilizamos o "método do cotovelo" (Robert L. Thorndike, 1953), um método que analisa o percentual de variância explicada em função do número de clusters. Por intermédio desse método, foi escolhido o número de agrupamentos de forma que a adição de outro agrupamento não apresenta grande importância para a modelagem dos dados e da solução.

Graficamente, o percentual de variância é explicado pelo número de grupos, os primeiros valores para a quantidade de grupos acrescentam muitas mudanças e informações, porém em um certo momento o ganho começa a cair, formando um ângulo no gráfico. O número de grupos é escolhido, neste ponto, por conseguinte, o "critério cotovelo". A percentagem de variância explicada é a razão entre a variação entre-grupo para a variância total, também conhecido como um teste de F.

Também é possível definir a quantidade de agrupamentos por intermédio do método citado utilizando a métrica SSE (Soma dos Erros Quadrados) (Draper, N.R.;

Smith, H., 1998; Pang-Ning Tan, *et al.*, 2005). Para cada ponto o erro é a distância para o agrupamento mais próximo. Para calcularmos o SSE, foi utilizado a Equação 15.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x) \tag{15}$$

Onde x é um ponto no agrupamento Ci e mi o ponto que representa o agrupamento Ci, comumente adotado o centróide. Dados dois agrupamentos, é escolhido o agrupamento com menor erro. Quanto maior o valor de *k*, menor o SSE.

A seguir na Figura 5 temos que o valor *k*=4 deve ser escolhido pelo "método do cotovelo" para nossa base de dados, no eixo horizontal representamos a variação da quantidade de agrupamentos e no eixo vertical a variação do valor de SSE.

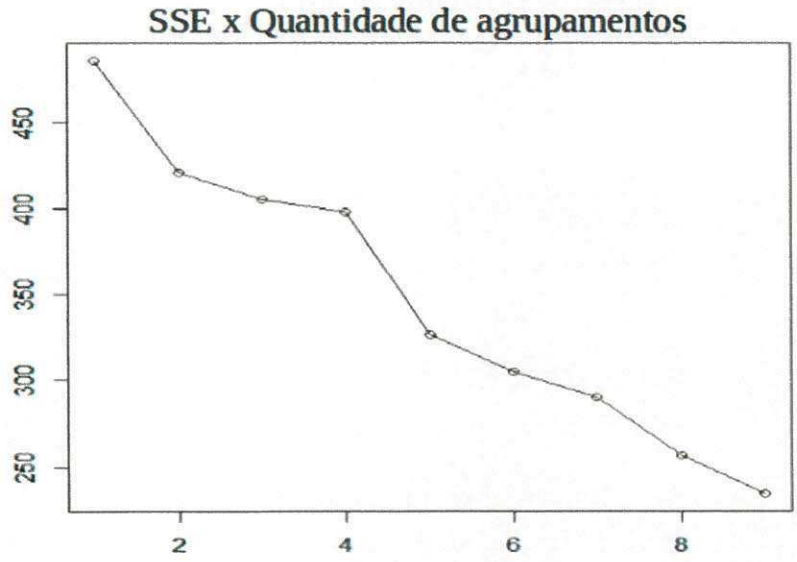

Figura 6 - Escolha do valor de k=4 pelo "método do cotovelo", utilizando calculando o SSE pela quantidade de agrupamentos.

Ao introduzir os vetores que representavam as tabelas acessadas por cada usuário e o valor de k igual a quatro no algoritmo K-means encontramos os agrupamentos que serão utilizados em nossa abordagem para recomendação descrita na seção a seguir.

5.3 Recomendações de consultas utilizando agrupamentos de usuários

O algoritmo de recomendação proposto tem como entrada o usuário que deseja receber as recomendações de consultas SQL (usuário alvo) e os agrupamentos gerados no processo descrito na seção anterior. Na primeira fase de nosso algoritmo, é identificado qual o grupo que o usuário pertence e selecionado este mesmo grupo para a próxima fase.

Na fase 2, por intermédio do algoritmo *k-Nearest Neighbors*, que calcula os k-usuários mais próximos (Altman, N. S., 1992) implementado pela ferramenta Weka, foram selecionados os k usuários mais próximos do usuário alvo, levando em consideração as tabelas que o usuário acessa por meio de suas consultas presentes no histórico de consultas de cada usuário. Nos estudos realizados foi definido empiricamente o valor de k igual a três.

Por fim, na fase 3, um programa desenvolvido em Java une os históricos de consultas dos três usuários selecionados, e de forma aleatória são escolhidas dez consultas que serão retornadas como recomendações para o usuário alvo.

Capítulo 6

Avaliação

Neste capítulo, é apresentado o método para avaliação e análise dos resultados encontrados após a execução dos experimentos realizados durante o estudo da abordagem proposta.

Na seção 6.1, é apresentada a base de dados utilizada neste trabalho de dissertação e o modo como ela foi utilizada nos estudos realizados. Na seção 5.2, é apresentado como as técnicas dos trabalhos selecionados para estudo, descritos na seção 3.2, foram avaliadas. Por fim, nas seções 6.3 e 6.4, são descritos os resultados obtidos juntamente com a análise realizada nesta pesquisa.

6.1 Base de dados utilizada

Foi utilizada nos estudos deste trabalho de dissertação uma base de dados disponibilizada pelo site do projeto SkyServer, constituída pelo histórico de consultas dos usuários do banco de dados do projeto durante o período de 03/2010 à 07/2011. Na base podemos observar o histórico de consultas provenientes de 447 endereços ips diferentes, neste trabalho de mestrado foram tratados cada um dos ips como um usuário, com média aproximada de 387 consultas por usuário, constituindo 79.538 consultas válidas que foram executadas pelos usuários.

6.2 Avaliações das técnicas selecionadas

Nesta seção, é detalhado o processo realizado para comparar o desempenho da abordagem proposta com o desempenho de técnicas já existentes, por meio das métricas relevância, quantidade média de novas tabelas, novidade e diversidade (detalhadas na seção 4.2).

Foi realizada a análise das recomendações de consultas após a aplicação das técnicas selecionadas para todos os usuários do histórico de consultas de banco de dados disponibilizado para nossa pesquisa pelo site SkyServer em duas etapas. Na primeira etapa da análise, foram calculadas as métricas de avaliação descritas no capítulo anterior para cada uma das técnicas selecionadas, utilizando todos os usuários presentes na base de dados deste estudo como entrada para as técnicas.

Ainda nessa 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. Os resultados encontrados a partir dessa etapa são apresentados na seção 6.3.

Na segunda etapa, os valores encontrados para cada uma das métricas foram comparados com o objetivo de gerar uma análise sobre o impacto da utilização de agrupamento de usuários para recomendação de consultas de banco de dados. Na seção 6.4, são apresentadas as análises realizadas nesta etapa.

6.3 Resultados obtidos

Após realizados os cálculos das métricas, conforme descritos no capítulo 4, a partir das recomendações realizadas pelas técnicas selecionadas para estudo deste trabalho, foram criadas as tabelas 4,5,6.

Cada técnica selecionada foi executada para recomendar consultas de banco de dados para todos os usuários da base de dados utilizada, separados em 4 grupos diferentes, agrupados segundo o método descrito na seção 4.1.3. Foram utilizados todos os históricos de consultas presentes na base (447 históricos) como entrada para cada algoritmo das técnicas selecionadas e as recomendações realizadas foram avaliadas de acordo com as métricas descritas na seção 4.2.

Tabela 4 - Métricas calculadas utilizando a técnica proposta por Arbanejad et al. (2010)

Arbanejad *et al. (2010)*						
Grupos	**Média Relevância**	**Média Tabelas Novas**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
1	0,14	1,02	0,27	0,31	1,17	14
2	0,1	1,30	0,18	0,27	1,63	42
3	0,14	1,68	0,24	0,51	1,00	19
4	0,09	0,93	0,15	0,19	0,81	17

Tabela 5- Métricas calculadas utilizando a técnica proposta por Stefanidis et al. (2009)

Stefanidis *et al. (2009)*						
Grupos	**Média Relevância**	**Média Tabelas Novas**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
1	0,05	2,75	0,12	0,11	0,78	9
2	0,06	2,83	0,08	0,21	1,38	16
3	0,26	1,8	0,4	0,9	0,22	5
4	0,26	1,67	0,41	0,79	0,2	5

Tabela 6- Métricas calculadas utilizando a técnica proposta por Chatzopoulou et al. (2011)

Chatzopoulou *et al. (2011)*						
Grupos	**Média Relevância**	**Média Tabelas Novas**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
1	0,33	0,06	0,73	0,64	1,39	20,00
2	0,25	0,11	0,53	0,49	1,74	40
3	0,29	0,2	0,6	0,57	1,32	19
4	0,45	0,07	0,96	0,89	0,71	9

Em seguida, foram realizadas recomendação de consultas de banco de dados utilizando a abordagem proposta por este trabalho de dissertação. Também foram feitas recomendações para todos os usuários da base de dados, separados nos 4 grupos

encontrados segundo o método para agrupamento descrito na seção 4.1.3.

Na abordagem proposta foram utilizadas informações sobre os grupos de usuários, dessa forma, foram feitas recomendações para usuários de cada grupo utilizando como entrada cada um dos grupos encontrados.

Por exemplo, usuários do grupo 1 foram utilizados para realizar recomendações para usuários dos grupos 1, 2, 3 e 4, e receberam recomendações geradas utilizando os grupos 2, 3 e 4. O mesmo foi feito para os demais grupos.

A seguir, nas tabelas 7,8,9 e 10 os resultados para as métricas estudadas ao utilizar a abordagem proposta nesta dissertação. Foram sinalizadas em amarelo os maiores valores encontrados para cada uma das métricas.

Tabela 7 - Utilizando como entrada os 4 grupos encontrados para realizar recomendações para o Grupo 1 de acordo com a abordagem proposta.

Grupo 1						
Grupos	**Média Tabelas Novas**	**Média Relevância**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
1	0,34	0,36	0,74	0,72	0,80	8
2	2,45	0,01	0,03	0,02	1,25	36
3	2,00	0,04	0,11	0,07	1,09	20
4	1,52	0,04	0,09	0,10	0,87	11

Para usuários pertencentes ao Grupo 1 são obtidos maiores valores para relevância, *precision* e *recall* utilizando usuários presentes no mesmo grupo que estes. Porém é possível observar que utilizando usuários do Grupo 2 são encontrados maiores valores para as métricas tabelas novas, novidade e diversidade.

Tabela 8 - Utilizando como entrada os 4 grupos encontrados para realizar recomendações para o Grupo 2 de acordo com a abordagem proposta.

Grupo 2						
Grupos	**Média Tabelas Novas**	**Média Relevância**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
1	2,60	0,02	0,03	0,06	1,18	18
2	1,08	0,06	0,13	0,13	1,37	25
3	1,84	0,03	0,06	0,09	1,47	27
4	1,58	0,04	0,07	0,09	1,33	18

Ao realizar recomendações para usuários do Grupo 2, novamente é possível verificar que as métricas relevância, *precision* e *recall* apresentam valores mais elevados quando são utilizados usuários do mesmo grupo dos usuários que recebem as recomendações. Porém desta vez para obter valores maiores para a métrica tabelas novas é necessário utilizar como entrada usuários do Grupo 1 e para as métricas novidade e diversidade é necessário utilizar usuários presentes no Grupo 3.

Nas tabelas 9 e 10 é observado o mesmo padrão descrito para o Grupo 1. Os valores para relevância, *precision* e *recall* são maiores utilizando usuários do mesmo grupo que os usuários alvos das recomendações, e são encontrados valores maiores para as tabelas novas, novidade e diversidade utilizando usuários do Grupo 2 como entrada para a abordagem descrita neste trabalho.

Tabela 9 - Utilizando como entrada os 4 grupos encontrados para realizar recomendações para o Grupo 3 de acordo com a abordagem proposta.

Grupo 3						
Grupos	**Média Tabelas Novas**	**Média Relevância**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
1	2,46	0,04	0,05	0,15	0,91	13
2	2,61	0,02	0,03	0,04	1,26	39
3	0,21	0,38	0,76	0,76	0,88	7
4	1,88	0,04	0,06	0,10	0,82	12

Tabela 10 - Utilizando como entrada os 4 grupos encontrados para realizar recomendações para o Grupo 4 de acordo com a abordagem proposta.

Grupo 4						
Grupos	**Média Tabelas Novas**	**Média Relevância**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
1	2,40	0,02	0,06	0,04	0,42	7
2	2,87	0,01	0,03	0,03	0,61	15
3	1,80	0,04	0,04	0,06	0,74	14
4	0,29	0,22	0,57	0,42	0,29	3

Ao fim da fase de obtenção dos valores das métricas estudadas, é possível agrupar os resultados encontrados das três técnicas estudadas e da abordagem proposta separados pelos grupos de usuários que receberam as recomendações, conforme apresentado nas tabelas 11 à 14. Os maiores valores encontrados utilizando a abordagem proposta (destacados em amarelo nas tabelas 7 à 10) foram utilizados para efetuar as comparações.

Tabela 11 - Comparando os valores das métricas encontrados para o Grupo 1 utilizando as técnicas selecionadas e a abordagem proposta.

Grupo 1						
Técnicas	**Média Tabelas Novas**	**Média Relevância**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
Arbanejad *et al. (2010)*	0,14	1,02	0,27	0,31	1,17	14
Stefanidis *et al. (2009)*	0,05	2,75	0,12	0,11	0,78	9
Chatzopoulou *et al. (2011)*	0,33	0,06	0,73	0,64	1,39	20,00
Abordagem proposta	2,45	0,36	0,74	0,72	1,25	36

Na tabela 11 é possível observar que a abordagem proposta neste trabalho de dissertação apresenta maiores valores para as médias das métricas tabelas novas, *precision, recall* e diversidade do que as demais técnicas.

Tabela 12 - Comparando os valores das métricas encontrados para o Grupo 2 utilizando as técnicas selecionadas e a abordagem proposta.

Grupo 2						
Técnicas	**Média Tabelas Novas**	**Média Relevância**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
Arbanejad *et al. (2010)*	0,1	1,30	0,18	0,27	1,63	42
Stefanidis *et al. (2009)*	0,06	2,83	0,08	0,21	1,38	16
Chatzopoulou *et al. (2011)*	0,25	0,11	0,53	0,49	1,74	40
Abordagem proposta	2,60	0,06	0,13	0,13	1,47	27

Para o Grupo 2, novamente é possível verificar que a métrica tabelas novas possui maior valor utilizando a abordagem proposta nessa dissertação. Esse fato também se repete nas tabelas 13 e 14, pois por meio da abordagem proposta é possível utilizar isoladamente usuários de grupos diferentes para recomendar consultas ao usuário alvo. As demais técnicas utilizam todos os usuários da base, assim podem ser selecionados usuários que acessaram a mesma tabela que o usuário alvo para realizar recomendações.

Tabela 13 - Comparando os valores das métricas encontrados para o Grupo 3 utilizando as técnicas selecionadas e a abordagem proposta.

Grupo 3						
Técnicas	**Média Tabelas Novas**	**Média Relevância**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
Arbanejad *et al. (2010)*	0,14	1,68	0,24	0,51	1,00	19
Stefanidis *et al. (2009)*	0,26	1,8	0,4	0,9	0,22	5
Chatzopoulou *et al. (2011)*	0,29	0,2	0,6	0,57	1,32	19
Abordagem proposta	2,61	0,38	0,76	0,76	1,26	39

Tabela 14 - Comparando os valores das métricas encontrados para o Grupo 4 utilizando as técnicas selecionadas e a abordagem proposta.

Grupo 4						
Técnicas	**Média Tabelas Novas**	**Média Relevância**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
Arbanejad *et al. (2010)*	0,09	0,93	0,15	0,19	0,81	17
Stefanidis *et al. (2009)*	0,26	1,67	0,41	0,79	0,2	5
Chatzopoulou *et al. (2011)*	0,45	0,07	0,96	0,89	0,71	9
Abordagem proposta	2,87	0,22	0,57	0,42	0,74	15

Ao fim da análise dos dados presentes nas tabelas 11 à 14, é possível verificar que por meio da abordagem proposta são encontrados maiores valores para as métricas estudadas em aproximadamente 58%, 70% e 66% dos casos, quando comparados com os valores encontrados utilizando as técnicas propostas por Arbanejad *et al.* (2010), Stefanidis *et al.* (2009) e Chatzopulou *et al.* (2011) respectivamente. Assim é possível afirmar que a abordagem proposta apresenta em média 64,6% dos casos valores maiores para a métricas observadas em relação as técnicas comparadas.

Na tabela 15 é observada a porcentagem de vezes que a abordagem proposta apresenta valores maiores para cada uma das métricas separadamente.

Tabela 15 - Comparando separadamente os valores das métricas encontrados utilizando as técnicas selecionadas e a abordagem proposta.

Abordagem proposta						
Técnicas	**Média Tabelas Novas**	**Média Relevância**	**Média Precision**	**Média Recall**	**Média Novidade**	**Média Diversidade**
Arbanejad *et al.* (2010)	100%	0%	75%	75%	50%	50%
Stefanidis *et al.* (2009)	100%	0%	75%	25%	100%	75%
Chatzopoulou *et al.* (2011)	100%	75%	50%	50%	25%	75%

Dessa maneira, é possível verificar que a abordagem proposta apenas não apresenta melhoria para os valores da métrica relevância quando são utilizadas as técnicas propostas nos trabalhos de Arbanejad *et al.* (2010) e Stefanidis *et al.* (2009) para realizar recomendações de consultas. Para as demais métricas estudadas, a abordagem proposta apresenta ganhos em até 100% dos casos considerando cada métrica de forma isolada.

Na seção 6.4, serão apresentadas as análises realizadas a partir dos resultados encontrados, bem como os passos aplicados para entender a razão do Grupo 2 ser o melhor grupo para se obter maiores valores para as métricas tabelas novas, novidade e diversidade no momento da recomendação para os demais grupos.

6.4 Análise dos resultados

Para análise dos resultados, os usuários de cada grupo encontrado por meio do algoritmo de agrupamento utilizado foram analisados utilizando os atributos armazenados dos históricos de consultas dos usuários, conforme descritas na seção 4.1.2. São eles: quantidade de horas no sistema, média de consultas realizadas por hora, quantidade de condições e tabelas acessadas por consulta.

Para facilitar a compreensão do comportamento dos usuários foram criados rótulos referentes ao perfil de comportamento observado dos usuários referentes as médias dos valores de cada uma das informações observadas. Em seguida as médias foram comparadas com as médias de outros usuários. Na tabela a seguir, são apresentados os resultados dessa comparação.

Tabela 16 - Comparação entre grupos de usuários de acordo com os atributos utilizados para gerar os agrupamentos.

Média de consultas realizadas por hora	Média de condições	Média de horas no sistema	Média de tabelas acessadas	Perfil de comportamento dos usuários
Menor	Maior	Maior	Menor	Especialistas
Menor	Menor	Menor	Maior	Pontuais
Maior	Menor	Maior	Menor	Iniciantes
Maior	Menor	Menor	Maior	Exploradores

As células da tabela 16 foram preenchidas com as palavras "Maior" e "Menor" quando os valores da média de cada uma das informações de cada usuário são maiores ou menores, respectivamente, que a média da informação considerando todosos usuários.

Assim, foram dados rótulos para identificar cada usuário da base de dados utilizada com o objetivo de facilitar o estudo e realizar referências aos comportamentos identificados. A seguir os nomes dos rótulos utilizados para descrever os perfis de comportamento encontrados dos usuários em contado com o banco de dados do projeto Skyserver, utilizado nesta pesquisa. Foram utilizadas as palavras "pouco(a)" e "muito(a)" para sinalizar quantidade abaixo ou acima da média dos outros usuários.

1. Especialistas: usuários que por um período acima da média utiliza o sistema de banco de dados na Web, mas que realizam poucas consultas com poucas tabelas em cada consulta e muitas condições. É assumido neste trabalho que os especialistas conhecem muito o tema do banco de dados e as condições que podem ser utilizadas para operar sobre os dados, porém possuem dificuldade com SQL e como formular consultas de banco de dados, por isso poucas consultas são realizadas. Também é possível observar que os Especialistas acessam poucas tabelas, demostrando interesse particular para estudo de partes do esquema do banco de dados;

2. Pontuais: usuários que acessam o sistema de banco de dados na Web por pouco tempo, realiza poucas consultas com poucas condições (consultas mais simples) em cada para acessar dados de muitas tabelas;

3. Iniciantes: usuários que utilizam o sistema de banco de dados na Web por um tempo acima da média em do sistema, porém acessam poucas tabelas por meio de muitas consultas simples, com poucas condições; e

4. Exploradores: usuários que permanecem por pouco tempo no sistema, porém executam muitas consultas com poucas condições, que retornam dados de várias tabelas do esquema.

Após essa análise dos comportamentos dos grupos, é verificado que o Grupo 2 utilizado para realizar recomendações de consultas apresenta maior quantidade de usuários do perfil Exploradores, o que justifica maiores valores para as métricas tabelas novas, novidade e diversidade.

Também é visto que quanto maior a quantidade de usuários do tipo Especialistas nos grupos, maiores serão os valores das métricas relevância, *precision* e *recall* quando usuários desse grupo forem usados para gerar recomendações para usuários do mesmo grupo.

Capítulo 7

Conclusões

7.1 Contribuições

Neste trabalho foram adaptadas as métricas relevância, novidade, diversidade e definida a métrica quantidade de tabelas novas para o problema de recomendação de consultas SQL em banco de dados e propomos uma nova estratégia para resolvê-lo. A estratégia utiliza agrupamentos dos usuários do banco de dados, elege quais agrupamentos apresentam maior valor para cada uma das métricas estudadas e seleciona consultas para recomendação desses agrupamentos.

Além disso, foram identificados quatro perfis de comportamento de usuários ao acessar um banco de dados na Web, por meio do estudo das informações: quantidade de horas no sistema, média de consultas realizadas por hora, quantidade de condições e tabelas acessadas por consulta, também definidas nesta dissertação.

Nossa estratégia aumenta os valores de métricas estudadas extraídas das recomendações em 64,6% dos casos na média quando comparada com técnicas presentes no estado-da-arte.

Assim, após a realização desta pesquisa, foi possível colaborar com estudos em recomendação de consultas SQL em banco de dados nos seguintes pontos:

1. Foram identificados perfis diferentes de usuários no sistema estudado e foi proposto um método para encontrar esses perfis. Dessa maneira, as hipóteses H1 e H2 foram aceitas.

2. Foram propostas e adaptadas métricas para avaliar a qualidade de recomendações de consultas de banco de dados. Assim, foi demonstrado que a hipótese H3 é verdadeira.

3. É possível recomendar consultas de um grupo específico de usuários, o que melhora a qualidade das recomendações, considerando as métricas estudadas. Assim, as hipóteses H4 e H5 foram aceitas.

4. Diferente de outros trabalhos, na abordagem proposta não são necessárias etapas para obtenção de preferências, o que agiliza o processo para realização de recomendações de consultas SQL em banco de dados.

5. Também não é exigido que o usuário do sistema tenha o conhecimento de todo o esquema do banco de dados.

6. Por meio da abordagem proposta é possível que novos usuários do sistema também recebam recomendações, uma vez que estes usuários podem ser considerados pertencentes ao grupo mais volumoso encontrado. Quando o sistema for atualizar os agrupamentos, os usuários novatos já terão históricos de consultas armazenados, serão colocados no grupo correto e terão seus perfis de comportamento identificados por meio das consultas realizadas.

7. Algumas técnicas para recomendação encontradas na revisão da literatura sofrem alterações quando expostas a constantes mudanças nos dados armazenados no banco de dados. Porém na abordagem proposta não existem essas alterações, uma vez que são considerados apenas os históricos de consultas dos usuários e não os dados retornados pelas consultas.

Esta pesquisa de dissertação também originou uma publicação no XII Workshop de Teses e Dissertações em Banco de Dados (WTDBD) do 28º Simpósio Brasileiro de Banco de Dados (SBBD), em 2013, com o título **Recomendação de Consultas de Banco de Dados utilizando Agrupamentos de Usuários**.

7.2 Trabalhos Futuros

Trabalhos futuros incluem a exploração de novas métricas para avaliação da qualidade das recomendações de consultas, uma vez que é possível extrair mais informações de consultas SQL, como por exemplo as colunas mais visualizadas das tabelas e informações presentes em outras cláusulas como HAVING, GROUP BY e ORDER BY. Além disso, poderiam ser propostos novos atributos para geração dos

agrupamentos.

Também são necessários experimentos de avaliação junto a usuários reais, tornando possível identificar quais são os melhores valores para as métricas estudadas, pois usuários reais podem preferir valores mais altos ou mais baixos dependendo da métrica, o que influencia na escolha de um grupo para realizar as recomendações.

Outro possível trabalho futuro, seria a realização de mais análises estatísticas sobre os dados e as métricas estudadas. Poderiam também ser realizados estudos sobre novas técnicas de agrupamento e novas técnicas para recomendação utilizando agrupamentos de usuários como, por exemplo, uma técnica composta pela união da abordagem proposta com técnicas presentes na literatura.

Por fim, em trabalhos futuros, poderia ser estudada a combinação das métricas apresentadas para realizar a recomendação conforme a abordagem proposta, e definir uma ordem para as consultas retornadas.

Referências Bibliográficas

AKBARNEJAD, J. *et al.* SQL querie recommendations. **In Proceedings of the VLDB Endowment**. v.3, n.1-2. p. 1597-1600, 2010

ALIGON, J., *et al.* Similarity measures for olap sessions. **In International Journal of Knowledge and Information Systems (KAIS)**. v.39, n.2, p. 463-489, 2014

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **The American Statistician.** v.46, n.3: p. 175–185, 1992

ARMSTRONG, J. S. Principles of Forecasting: A Handbook for Researchers and Practitioners. **International Series in Operations Research & Management Science.** v.30, 2001.

BACKER, E. Computer Assisted Reasoning in Cluster Analysis. **Prentice Hall.** V.1, 1995.

SIGURBJÖRNSSON, B.; VAN ZWOL, R. Flickr tag recommendation based on collective knowledge. **In Proceedings of the 17th International Conference on World Wide Web (WWW'08)**. p.327-336, 2008.

CHATZOPOULOU, G., *et al.* The QueRIE system for personalized query recommendations, IEEE Data Eng. Bull. v.34(2), p.55–60, 2011

DRAPER, N.R.; SMITH, H. Applied Regression Analysis (3rd ed.). 1998

BELÉM, F., *et al.* Explorando Relevância, Novidade e Diversidade em Recomendação de Tags. **In Anais do Simpósio Brasileiro de Sistemas Multimídia e Web**. 2012.

CHATZOPOULOU, G., *et al.* Query recommendations for interactive database exploration. **In Scientific and Statistical Database Management Lecture Notes in Computer Science.** v. 5566, p. 3–18, 2009.

LINDEN, G., *et al.* Amazon.com Recommendations: Item-to-Item Collaborative Filtering. **In Internet Computing, IEEE**. v.7, n.1, p.76-80, 2003.

PANG-NING TAN, *et al.* Introduction to Data Mining. Edition: 1, 2005

JAIN, A.K.; DUBES, R.C. Algorithms for Clustering Data. **Prentice Hall Advanced Reference Series: Computer Science**. 1988

KHOUSSAINOVA, N., *et al.* Snipsuggest: context-aware autocompletion for SQL. **In Proceedings of the VLDB Endowment**. v.4 n.1, p. 22-33. 2010

KOUTRIKA, G.; IOANNIDIS, Y. Personalization of Queries in Database systems. **In**

Proceedings of 20th Intl. Conf. On Data Engineering (ICDE). p. 597-608. 2004

LIMA, F., *et al.* Revisitando Técnicas de Bancos de Dados no Contexto da Web. **In: PUC-RIO**. 1999.

LIMAM, L., *et al.* 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. 2010

MAES, Pattie; SHARDANAND, Upendra. Social Information Filtering: Algorithms for Automating "Word of Mouth". **In CHI '95 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems**. p. 210-217, 1995.

MACQUEEN, J. B. Some Methods for classification and Analysis of Multivariate Observations. **In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.** p. 281–297, 1967

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

MURTHI, B. P. S.; SARKAR, S. The role of the management sciences in research on personalization. **In Manage Science.** V. 49, n. 10, p. 1344-1362, 2003

KHOUSSAINOVA, N., *et al.* A case for a collaborative query management system. **In 4th Biennial Conference on Innovative Data Systems Research (CIDR)**. 2009.

ELMASRI, R.; NAVATHE, S. B. Sistemas de Banco de Dados 6.ed. **Pearson Addison Wesley**. 2011.

RESNICK, P.; VARIAN, H. R. Recommender Systems. **In Communications of the ACM.** v. 40, p. 56-58, 1997.

THORNDIKE, R. L. Who Belong in the Family?. **In Psychometrika**. v.18, n.4, p 267–276, 1953.

MCNEE, S., et al. Being accurate is not enough: how accuracy metrics have hurt recommender systems. **In Proceeding CHI EA '06 CHI '06 Extended Abstracts on Human Factors in Computing Systems.** p. 1097-1101, 2006.

VARGAS, S.; CASTELLS, P. Rank and Relevance in novelty and diversity metrics for recommender systems. **In Proceeding RecSys '11 Proceedings of the fifth ACM conference on Recommender systems.** p. 109-116, 2011.

SARKAS, N., et al. Measure-driven keyword-query expansion. In Proceedings of the VLDB Endowment. v. 2, n. 1, p. 121-132, 2009.

SARMA, A.D., et al. Synthesizing view definitions from data. **In Proceeding ICDT '10 Proceedings of the 13th International Conference on Database Theory**. p. 89-103, 2010.

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

Site do projeto Skyserver disponível em http://cas.sdss.org/ - acessado em 11/04/2014

Site do software Java disponível em https://www.java.com/pt_BR/- acessado em 14/07/2014

Site da ferramenta de desenvolvimento Netbeans disponível em https://netbeans.org/ - acessado em 14/07/2014

Site da ferramenta Weka disponível em http://www.cs.waikato.ac.nz/ml/weka/ - acessado em 14/07/2014

STEFANIDIS, K., *et al.* You May Also Like results in relational databases. **In Proc. PersDB**. p. 37-42, 2009.

TRAN, Q.T.; CHAN, C.Y. How to conquer why-not questions. **In SIGMOD '10 Proceedings of the 2010 ACM SIGMOD International Conference on Management of data.** p. 15-26, 2010.

YANG, X., *et al.* Recommending join queries via query log analysis. **In 25th International Conference on Data Engineering (ICDE 2009)**. p. 964–975, 2009.

ZHANG, Z.; NASRAOUI, O. Mining search engine query logs for query recommendation, **In Proceedings of the 15th international conference on World Wide Web**. p. 1039-1040, 2006