ESTRATÉGIAS DE ORDENAÇÃO PARA COMPLETAR CONSULTAS EM E-COMMERCE

VICTOR COSTA OLIVEIRA

ESTRATÉGIAS DE ORDENAÇÃO PARA COMPLETAR CONSULTAS EM E-COMMERCE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Edleno Silva de Moura

Manaus

Agosto de 2018

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Oliveira, Victor Costa de
O48e Estratégias de Ordenação para Completar Consultas em E-Commerce / Victor Costa de Oliveira. 2018
74 f.: il. color; 31 cm.

Orientador: Edleno Silva de Moura
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. autocomplete. 2. consultas. 3. ecommerce. 4. ordenacao. I. Moura, Edleno Silva de II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

" Estratégias de Ordenação para Completar Consultas em E-Commerce"

VICTOR COSTA DE OLIVEIRA

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Edleno Silva de Moura - PRESIDENTE

Prof. André Luiz da Costa Carvalho - MEMBRO EXTERNO

Prof. Leandro Balby Marinho - MEMBRO EXTERNO

Manaus, 28 de Agosto de 2018

À minha esposa, a quem amo incondicionalmente.

Agradecimentos

À Deus, Autor da Vida, Salvador e Consolador. A Ele toda a Glória!

À minha esposa, Kimberly Graziela, que caminha a jornada da vida ao meu lado e compartilha todos os momentos, uns maravilhosos e outros bem difíceis. Estaremos juntos nessa vida e após ela também.

Ao Nicholas e Laura, que já existem em nossos corações.

Aos meus pais, José Fernandes e Valdenízia, que me criaram e educaram da melhor forma possível.

Aos meus irmãos, Selma, Joice e Jônatas, pelo cuidado e carinho de verdadeiros irmãos.

Aos meus sobrinhos e cunhados pela alegria que trouxeram a esta família.

À minha segunda família, Amorim, pelo cuidado e apoio como se fora família de sangue.

Ao meu pastor, Raimundo Feitozas, toda sua família e os irmãos de fé da Igreja Presbiteriana do Coroado III pelas orações, carinho, compreensão, incentivo em todos os momentos e companheirismo digno de uma família cristã.

Aos meus amigos tão chegados quanto irmãos que com suas amizades tornam o dia a dia mais simples e divertido.

Aos meus companheiros de trabalho da Linx que permitiram e incentivaram o estudo e a dedicação a este trabalho, entendendo quando precisei concentrar esforços neste e também por ouvirem o andamento de cada decisão.

Ao meu professor e orientador, Edleno, por ter sido o guia deste trabalho, vendo alternativas muito além das palavras e gráficos, e mostrando o caminho correto a seguir.

"Pois o Senhor é quem dá sabedoria; e da sua boca procedem o conhecimento e a inteligência."
(Salomão)

Resumo

O número de pessoas que compram produtos em sites de comércio eletrônico vem crescendo nos últimos anos. A facilidade e conforto aliado à crescente segurança dos sites possibilitou que mais pessoas comprem produtos com apenas alguns cliques. Neste universo, um dos primeiros sistemas que um usuário se depara no processo de compra é o Complemento Automático de Consultas, que sugere consultas e produtos a cada letra digitada na caixa de busca pelo comprador. Saber quais consultas sugerir a cada letra digitada e ordená-las da melhor forma possível é o desafio deste sistema. Este trabalho se concentra em saber qual é a melhor fonte de geração dessas consultas e estuda estratégias de ordenação destas no cenário do E-Commerce. Em toda nossa pesquisa não achamos um trabalho que estuda esse tema neste cenário específico, fazendo este trabalho ser o pioneiro no assunto. As estratégias implementadas neste trabalho foram testadas em lojas reais do varejo online brasileiro e latino-americano e os resultados foram diferentes em cada loja. O MRR em uma loja brasileira de cosméticos, por exemplo, usando uma das estratégias deste trabalho, atingiu quase 0,8 para prefixos com três letras digitadas e 0,6 para uma loja de Eletrodomésticos com o mesmo tamanho de prefixo.

Abstract

The number of people who buy products on e-commerce sites has been growing in recent years. The ease and comfort coupled with the growing security of the websites made it possible for more people to buy products with just a few clicks. In this universe, one of the first systems that a user encounters in the purchase process is the Automatic Query Complement, which suggests queries and products for each letter entered in the search box by the buyer. Knowing which queries to suggest to each letter typed and sorting them in the best possible way is the challenge of this system. This work focuses on knowing the best source of these queries and examines strategies for sorting queries in E-Commerce. In all of our research we did not find a paper that studies this theme in this scenario, making this work the pioneer in the subject. The strategies implemented in this paper were tested in Brazilian and Latin American retail real stores. The results were different in each store. The MRR at a Brazilian cosmetics store, for example, using one of the strategies developed in this paper, reached almost 0.8 for prefixes with three typed letters and 0.6 for an Appliances store with the same prefix size.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Contexto Geral

O número de consumidores do comércio eletrônico no Brasil passou de um milhão em 2001 para quase 60 milhões em 2014 (ver Figura 1.1, fazendo com que os lojistas dessem mais atenção ao público que está migrando da loja física para a virtual. Os sistemas de buscas dentro de sites de comércio eletrônico, portanto, tornaram-se ferramentas das mais importantes no processo de compra.

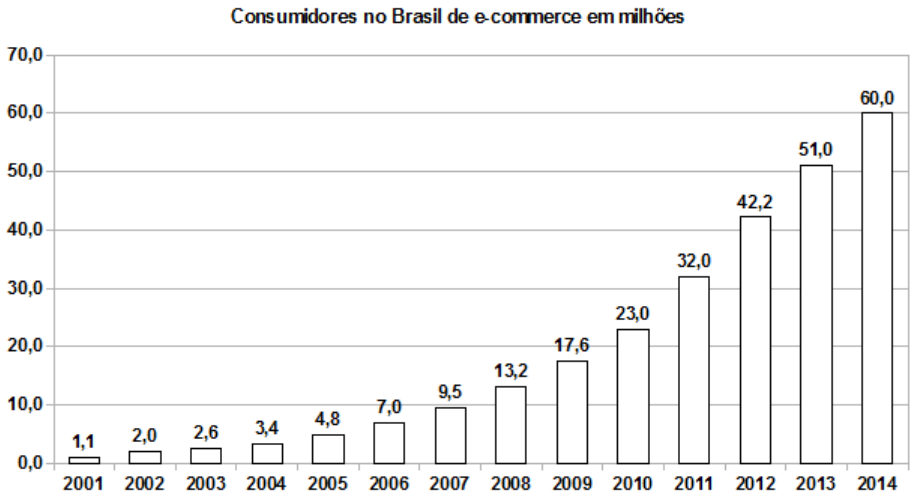


Figura 1.1. Crescimento do número de consumidores no Brasil de 2009 a 2013

As lojas virtuais, também chamadas de comércio eletrônico, representam um novo modelo procurado por pessoas que preferem o conforto de sua casa para comparar preços e realizar compras. À medida que os sistemas de entrega das lojas virtuais estão se tornando mais rápidos e confiáveis, o número de interessados em comprar sem sair de

casa vem aumentando. Apesar disso, o número ainda é pequeno em relação ao varejo físico, o que deixa a impressão que ainda haverá grandes aumentos e, consequentemente, grandes demandas para os sistemas de busca.

Segundo dados obtidos pelo *Centre for Retail Reasearch* do Natal do ano de 2015 (ver Figura 1.2), o Brasil movimentou um montante de 2,14 bilhões de euros por meio de lojas de comércio eletrônico. Aparentemente um valor expressivo, mas quando comparado à Grã-Bretanha, EUA e alguns países da Europa, esse valor ainda é bastante baixo, mostrando que nesses países, o ato de comprar online tem potencial para uma participação bem maior em nossa economia.

Figura 1.2. Quantidade de vendas no Natal de lojas de comércio eletrônico comparadas com o varejo físico.

Um dos grandes desafios de lojas de comércio eletrônico é desenvolver um sistema de busca robusto e inteligente o suficiente para atender usuário de forma assertiva e eficiente. O atendimento ao cliente e o entendimento de suas necessidades são fatores determinantes no sucesso de um empreendimento. Esse atendimento ao usuário final ganha muito mais importância quando vamos para o varejo online. Ali, os clientes tendem a ser mais exigentes: bastam alguns cliques para mudarem de loja. Portanto, é imprescindível criar motivos para envolver o visitante com ferramentas úteis para auxiliá-lo a achar o produto desejado. Quanto melhor o sistema de buscas utilizado, maior a chance de o usuário encontrar o que procura e ficar satisfeito com a experiência de compra.

Podemos dizer que o sistema de busca nas lojas virtuais tem papel semelhante aos atendentes do varejo físico. Tudo o que o cliente quiser saber sobre determinado produto deve ser atendido e respondido, coerentemente, por esse sistema. Ele é respon-

sável por direcionar os usuários a verem os produtos de seu interesse, seja mostrando resultados a partir de cliques no menu de navegação, seja mostrando páginas de ofertas exclusivas criadas pelo lojista ou ainda exibindo produtos para uma consulta feita no campo de busca. Na Figura 1.3 temos exemplos de onde o sistema de busca atua na página principal de um site.

Figura 1.3. Diferentes formas de interação com o Sistema de Busca. (1) O usuário pode digitar uma consulta na caixa de busca. Em (2) o usuário pode navegar pelo Menu de opções. (3) representa links para buscas que o lojista posta na sua vitrine.

Um dos componentes mais importantes em um Sistema de Busca é conhecido como Sistema de Complemento Automático de Consultas (do inglês, Query Auto Completion), o qual doravante será citado como SCAC ao longo deste trabalho. O SCAC sugere algo que o cliente já pensou, mas ainda não terminou de digitar. Ele ajuda usuários a expressarem o que desejam, completando as consultas que começaram a digitar e deixando o processo de busca mais simples e rápido. Ele também auxilia na ortografia da consulta, sugerindo correções de termos durante o processo de digitação, seja por erros ortográficos ou por falhas de digitação.

O SCAC tem se tornado cada vez mais importante dentro de sites em geral, sobretudo em lojas virtuais, já que ele é uma das primeiras interações desses usuários com o site de vendas. É através dele que o cliente começa a conhecer os itens que estão à venda na loja.

A porcentagem de aumento de vendas nas lojas que passaram a usar o SCAC varia de acordo com as características de cada uma. Segundo John [6], vemos que a porcentagem de vendas diretas, em média, aumenta de 2% a 8%. Fala-se vendas diretas porque não se pode calcular com precisão o quanto um SCAC pode acrescentar no faturamento por vendas futuras de usuários que voltaram ao site por sempre acharem o que desejam.

1.2 Problema abordado

Saber quais consultas sugerir para o usuário no momento em que o mesmo as digita é uma tarefa de fundamental importância, haja vista que o sistema deve considerar várias informações em um curto intervalo de tempo, informações estas que são oriundas de diversas fontes: do catálogo de produtos da loja e dos próprios usuários que utilizam e acabam realimentando o sistema com suas interações.

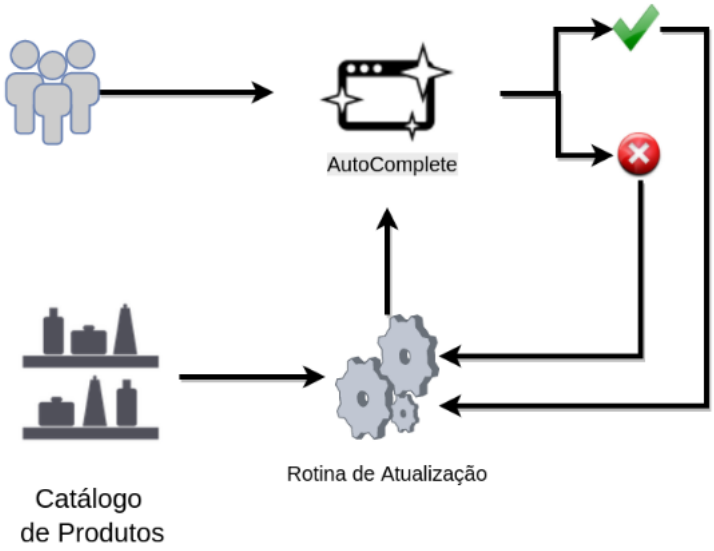


Figura 1.4. Arquitetura do Sistema de SCAC mostrando as interações dos usuários e do catálogo de produtos e como são usados para gerar e melhorar as sugestões do SCAC.

O objetivo deste trabalho vai além de saber quais consultas sugerir ao cliente. Apresentamos aqui diferentes estratégias de extração de características e ordenação dessas consultas no intento de fazer com que as sugestões mostradas sejam as melhores possíveis, uma vez que elas aumentam a chance de o usuário interagir mais com o sistema e de realizar a compra. O conceito de "melhores possíveis"é bem subjetivo e, por isso, estudamos algumas formas de avaliar e quantificar os resultados obtidos. Como exemplo da dificuldade que se pode ter em avaliar um SCAC, é possível que um conjunto de buscas aumente o faturamento da loja, mas diminua a quantidade de cliques em produtos da mesma.

As estratégias desenvolvidas neste trabalho para aprimorar resultados obtidos por SCACs variam na escolha do conjunto de características e informações a serem utilizadas, na forma de calcular a ordem das consultas e no objetivo dessa ordenação.

Com base no objetivo delineado acima, realizamos um estudo sobre quais informações são importantes para serem consideradas no cálculo da ordem mostrada ao

usuário. Classificamos as informações em grupos de acordo com a fonte desses dados, e avaliamos os resultados de diferentes algoritmos propostos na literatura para calcular a importância de cada consultas candidata a ser mostrada como sugestão em um SCAC. Além disso, verificamos, em algumas lojas do comércio eletrônico, a diferença do comportamento dos usuários quando eles são expostos a diferentes tipos de ordenação de consultas.

Até o momento da publicação deste trabalho, não encontramos nenhuma pesquisa em que se tratasse métodos para ordenação de sugestões de busca especificamente para o comércio eletrônico. O estudo de SCAC na literatura foi limitado ao contexto de sistemas buscas genéricas para a Web, tais como o Google (http://www.google.com) e Bing (http://www.bing.com), onde o objeto a ser buscado pode ser qualquer tipo de conteúdo disponível na Web, incluindo informações, páginas, imagens e demais tipos de conteúdo disponíveis. Assim, este trabalho busca ser o pioneiro no estudo de diferentes métodos de ordenação para complementar consultas em lojas de comércio eletrônico.

1.3 Hipóteses a serem validadas

O contexto do comércio eletrônico é diferente do mundo de buscas na Web. As buscas nos sites de comércio eletrônico tendem a serem voltadas majoritariamente ao processo de compra de um produto, enquanto em buscas na Web os usuários podem ter diversos outros propósitos. Em segundo lugar, os objetos de busca são diferentes, visto que que websites têm características bem diferentes dos produtos. Os produtos têm informações estruturadas, tais como categoria, atributos, preço e data de lançamento. Os sites, por sua vez, não têm estrutura definida, são bastante diferentes entre si e seu conteúdo é formado por textos, imagens, vídeos, tabelas e diversos outros tipos de informação. Essas especificidades fazem com que SCACs para ambientes de comércio eletrônico possam utilizar informação diferente e ter comportamentos diferentes dos desenvolvidos para a Web, justificando a realização de um estudo específico para desenvolver SCACs para comércio eletrônico.

Nesta dissertação, estudamos os métodos de ordenação de sugestões de consultas já conhecidos na literatura desenvolvidos para o ambiente de buscas na Web e utilizamos em ambientes de comércio eletrônico para medir o impacto destes nesse novo cenário. Além disso, estudamos novas características e desenvolvemos outros métodos de ordenação com a finalidade de verificar qual deles apresenta os melhores resultados em diferentes segmentos do varejo. Para tanto, avaliamos cada método em lojas de diferentes seguimentos com o objetivo de verificar qual método se destaca em cada

cenário.

Em primeiro lugar, verificou-se a importância de cada fonte de geração de consultas para o contexto de SCAC no E-Commerce, comparando a taxa de visualização e cliques nas consultas geradas de cada fonte para algumas lojas do E-Commerce brasileiro. Em segundo lugar, comparou-se o método utilizado na literatura para ordenação de consultas com novos métodos desenolvidos neste trabalho utilizando os conceitos do ambiente de comércio eletrônico. Por último encontrou-se a relação entre qualidade dos métodos de ordenação com a quantidade de dias que se deve extrair de informação para estes métodos.

O trabalho está dividido como segue: no Capítulo 2 falaremos dos Conceitos Básicos e Trabalhos Relacionados; no Capítulo 3 apresentamos as Estratégias de Ordenação Propostas; no Capítulo 4 explicaremos os Experimentos Realizados e os Resultados Obtidos e, no Capítulo 5 ,traremos as conclusões sobre o trabalho.

Capítulo 2

Conceitos Básicos e Trabalhos Relacionados

Neste capítulo entenderemos os conceitos relacionados com o problema de ordenação do SCAC assim como os trabalhos cujos assuntos são próximos ao tratado nesta dissertação. É importante salientar, logo no início, que, em todo nosso estudo e pesquisa, não encontramos nenhum trabalho estritamente ligado com nosso tema, ou seja, com SCAC para lojas de comércio eletrônico. Os trabalhos encontrados e tratados neste capítulo se relacionam por tratar de estudos para o desenvolvimento de SCACs, mas no contexto de Máquina de Busca para WEB. Esses dois ambientes possuem características bastante distintas entre si.

A construção do Sistema de SCAC para comércio eletrônico levou em consideração dois trabalhos chave. Em Chaudhuri & Kaushik [5] vimos como construir o algoritmo capaz de receber um prefixo como entrada e sugerir sugestões de busca com tolerância a erro de edição e em Bar-Yossef & Kraus [1] aprendemos primeiramente como e de onde extrair as sugestões de busca e, em segundo lugar, como utilizar informações sobre a frequência das consultas em diferentes intervalos de tempo na ordenação das sugestões.

O capítulo divide-se em três grupos: Geração de Consultas, Processamento do Prefixo e Pontuações e Ordenação. O Sistema de SCAC é composto, principalmente, dessas três etapas e, por causa disso, estudamos os artigos que são estado-da-arte em cada uma delas.

2.1 Geração de Consultas

Em qualquer Sistema de Busca - como é o caso de SCACs - os objetos a serem buscados já são conhecidos de antemão. Em sistemas de biblioteca, por exemplo, os livros são

catalogados e, assim, uma busca por um autor qualquer deve trazer somente os livros desse autor que estão registrados no sistema, não sendo possível trazer outros livros "de fora"desse ambiente, mesmo que esse autor os tenha escritos. Semelhantemente, uma máquina de buscas para a Web só traz como resultados sites em que ela própria já visitou e catalogou anteriormente. Embora óbvio, esse conceito é de supra importância para o SCAC. As consultas que serão sugeridas precisam ser pré-computadas anteriormente, em um processo que chamamos de Geração de Consultas.

Cai et al. [4] demonstra uma arquitetura (ver Figura 2.1) de um SCAC em que as sugestões são extraídas do Log de Consultas dos próprios usuários que utilizam o sistema, ou seja, um sistema retro-alimentado. Bar-Yossef & Kraus [1] e Mitra & Craswell [9] também utilizam consultas dos usuários como entrada para o SCAC.

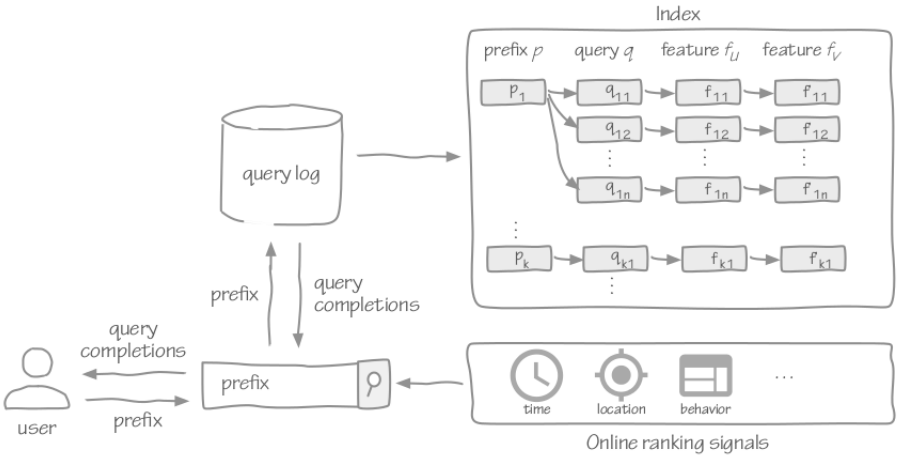


Figura 2.1. Arquitetura do Sistema de SCAC apresentado por Cai et al. [4]

O Log de Consultas pode ser representado como um banco de dados que armazena as consultas submetidas pelos usuários ao Sistema de Busca. Bar-Yossef & Kraus [1] afirma que "The search engine suggests to the user the completions that have been most popular among users in the past", ou seja, usuários tendem a repetir as mesmas consultas que outros usuários já fizeram no passado, fazendo do log de consultas uma fonte importantíssima para geração de consultas

Um caso diferente é tratado em Mitra & Craswell [9], onde é feito um algoritmo capaz de sugerir consultas pouco ou nunca antes requisitadas. Isso é possível quando utilizamos substrings não-raras para gerar sugestões em tempo real. Na Tabela 2.1 temos um exemplo de sugestões de busca para o prefixo *what to cook with chicken and broccoli and*. Nesse caso, é possível gerar consultas *sintéticas* utilizando não o prefixo inteiro como entrada, mas apenas alguns dos últimos termos para que seja possível criar

sugestões que completem o prefixo do usuário. No exemplo da Tabela 2.1, é possível que o prefixo requisitado para o SCAC seja apenas *broccoli and*. Certamente não usar o prefixo todo tem os efeitos colaterais na qualidade das sugestões, mas os benefícios trazidos os superam, segundo o autor.

Tabela 2.1. Exemplo de sugestões de consultas para prefixo raros mostrado por Mitra & Craswell [9]

what to cook with chicken and broccoli and
what to cook with chicken and broccoli and bacon
what to cook with chicken and broccoli and noodles
what to cook with chicken and broccoli and brown sugar
what to cook with chicken and broccoli and garlic
what to cook with chicken and broccoli and orange juice
what to cook with chicken and broccoli and beans
what to cook with chicken and broccoli and onions
what to cook with chicken and broccoli and ham soup

2.2 Processamento do Prefixo

O processamento do prefixo é a tarefa de encontrar consultas que complementam as letras até então digitadas por um usuário. O processo de obtenção de consultas candidatas à resposta dos prefixos digitados não é uma tarefa trivial. Verificar quais são essas sugestões e ordená-las com qualidade possui um custo associado que precisa ser levado em conta quando se tem que responder em um tempo curto. Ao priorizar o tempo de resposta, pode-se perder informações importantes para a qualidade da mesma. Em contrapartida, processar muitas características com ênfase na precisão pode ser custoso e fazer com que o sistema tenha seu desempenho piorado e, como vimos anteriormente, o SCAC precisa ser bem rápido para dar impressão de instantaneidade ao usuário.

A técnica mais comum para processar prefixos na literatura é a estrutura de dados chamada árvore de prefixos. A versão mais básica dessa árvore é aquela onde cada nó representa um caracter de uma consulta (Ver Figura 2.3). Chaudhuri & Kaushik [5] utilizou essa estrutura para, além de realizar a tarefa de gerar as sugestões a partir de um prefixo, também conseguir trazer sugestões mesmo que elas não tenham exatamente o prefixo digitado, ou seja, é possível obter sugestões mesmo com prefixos digitados incorretamente, utilizando o conceito de distância de edição de Levenshtein [7].

O processamento do prefixo em uma trie é feito da seguinte forma: Primeiramente, insere-se as sugestões pré-computadas nessa estrutura de árvore n-ária, onde cada letra

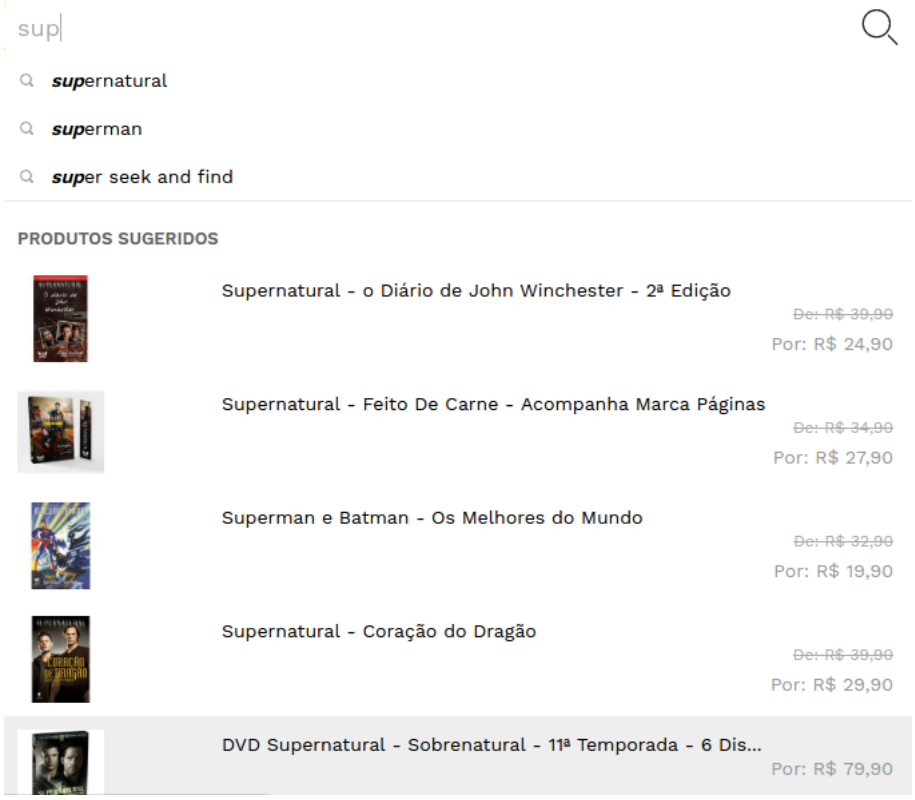


Figura 2.2. Exemplo de sugestões retornadas para o prefixo *sup* em uma Loja de Livraria Online

é um nó pai da letra seguinte. O nó filho da última letra é representado por um nó especial, simbolizado por # na Figura 2.3. Para obter as sugestões dado um prefixo p = *'th'*, o algoritmo de busca percorre a árvore visitando, a partir do nó raiz (*root*) os nós t e h, respectivamente, como pode ser visto na Figura 2.4. Ao chegar no nó h, as sugestões são todos os caminhos que levam a algum nó terminal, simbolizado por # e, no caso do algoritmo proposto por Chaudhuri & Kaushik [5], alguns caminhos *irmãos* de h, considerando um limite máximo de erros a ser tolerado. Nesse caso, retornaria *the (erro = 0), there (erro = 0), town (erro = 1) e tavern (erro = 1)*

Outras estruturas são encontradas na literatura como opção para obtenção de sugestões baseadas em prefixo em SCAC. É o caso de Li et al. [8] que propôs um método baseado em probabilidades. Cada sequência de letra representa um estado e probabilidades da próxima letra são realizadas para sugerir consultas.

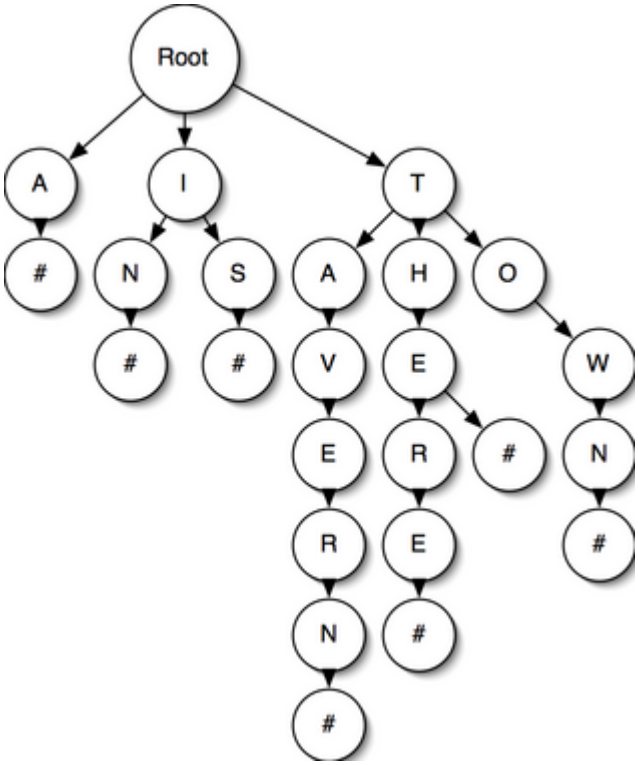


Figura 2.3. Exemplo de uma trie em funcionamento com os termos: a, in, is, tavern, there, the e town inseridos.

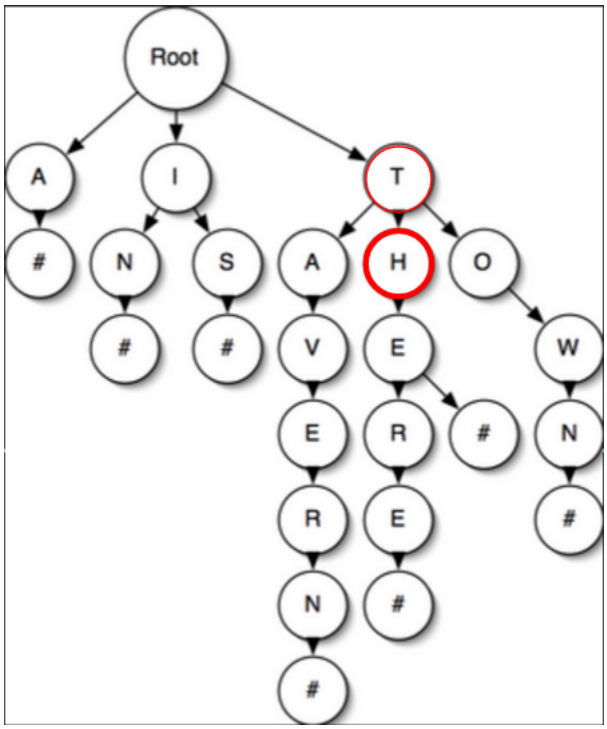


Figura 2.4. Exemplo de uma busca na trie pelo prefixo 'th'.

2.3 Pontuações e Ordenação

Uma vez que as sugestões que complementam o prefixo são obtidas, saber a ordem de cada uma é outra tarefa não-trivial de suma importância para o desempenho dos SCAC. É nessa etapa que a maioria dos trabalhos de SCAC tem seu enfoque [[11], [4], [1], [12]]. A quantidade de informação que pode ser utilizada nessa etapa é muito grande e variada. Shokouhi [11] aplicou *Machine Learning* para usar dados de gênero, idade e localização dos usuários para melhorar a ordem das sugestões, ou seja, essas características foram usadas para pontuar as sugestões e o *Machine Learning* foi usado para ordená-las. Bar-Yossef & Kraus [1] foi o primeiro a utilizar diferentes quantidades de log de consultas para pontuar e ordenar as sugestões e Whiting & Jose [12] conseguiu equilibrar a robustez de consultas muito frequentes do passado com dados de log com consultas submetidas recentemente ao sistema, gerando boas pontuações para sugestões que são tendências no momento.

O trabalho de Bar-Yossef & Kraus [1] utiliza a frequencia das consultas em dias passados para prever a pontuação das consultas para SCAC. Este será o baseline deste trabalho, onde compararemos este método no contexto do Comércio Eletrônico, juntamente com outros algoritmos que substituem a simples frequência passada para quantidade de vezes que uma consulta levou ao usuário clicar em um produto e a quantidade de vezes que uma consulta levou o usuário a comprar um produto.

Já o trabalho de Whiting & Jose [12] também foi utilizado por considerar diferentes quantidades de dias no passado para atribuir as pontuações para as consultas do SCAC, variando de 2, 4, 7, 14 e 28 dias. Neste trabalho utilizamos essa quantidade para verificar as diferentes pontuações que gerariam, acrescentando ainda a quantidade de 90 dias.

Métodos de Ordenação para uma lista de objetos são bastante estudados na literatura. As consultas em SCAC podem ser vistas como objetos que possuem várias características diferentes como, por exemplo, a quantidade de vezes que foi realizada nos últimos X dias, a quantidade de produtos que retornam, dentre outras coisas. Em diversos cenários, incluindo o SCAC, cada objeto possui dezenas de características diferentes, e saber como combiná-las é um desafio à parte. Técnicas de *Aprendizagem de Máquina* para descobrir o peso ideal de cada característica são bastante estudados no contexto de busca, mas pouco foi falado o assunto no contexto de SCAC. Shokouhi [11] utilizou métodos de *Aprendizagem de Máquina* para descobrir as configurações ideais e chegar em um modelo que consegue ter bons resultados. O algoritmo escolhido foi *Lambda-MART*, prosposto por Burges et al. [2] baseado em *Árvores de Decisões*.

Consultas relacionadas também foram estudadas na literatura. Cai & de Rijke

[3] propôs uma forma de não exibir duas ou mais consultas que queiram dizer a mesma coisa, ou seja, duas consultas sinônimas. Esse método foi aperfeiçoado nesta dissertação aplicando o conceito de consultas sinônimas para não apenas uma permutação de termos, a qual Cai & de Rijke [3] faz referência, mas também a consultas que diferem em número (plural-singular) e grau (masculino-feminino).

Em Rangel [10] percebemos a diferença entre os ambientes de uma loja física para online. Em lojas de comércio eletrônico, os usuários podem sair do site do lojista com um clique. Isso faz com que os sistemas de Busca e SCAC precisem ser cada vez mais precisos para não perder os clientes. Capturar as informações utilizadas pelos trabalhos relatados no contexto de comércio eletrônico também não é tarefa simples. No próximo capítulo veremos com mais detalhes os conceitos utilizados no sistema desenvolvido para as maiores lojas de comércio eletrônico da América Latina.

Capítulo 3

Estratégias de Ordenação para Completar Consultas em E-Commerce

Neste capítulo, propõe-se relatar os caminhos escolhidos para atingir os objetivos e validar as hipóteses mencionadas nesta dissertação (Ver Seção 1.3). As estratégias de ordenação escolhidas para completar consultas no comércio eletrônico passam pela escolha da fonte geradora de consultas, o processamento do prefixo e, principalmente, das escolhas das características das consultas e o peso de cada uma, além da escolha do algoritmo de ordenação das consultas.

Antes de entrar no mérito das estratégias de ordenação (falaremos com mais detalhes na seção 3.4), mostraremos a arquitetura geral do SCAC desenvolvido, as particularidades da Geração de Consultas e os detalhes do Processamento do Prefixo do Complemento Automático de Consultas.

3.1 Arquitetura do Complemento Automático de Consultas

O Sistema de Complemento Automático de Consultas (SCAC) desenvolvido é capaz de atingir bons desempenhos em nichos diferentes de lojas virtuais. A arquitetura foi construída com tecnologias de *BigData* e processamento em tempo real para garantir alta disponibilidade, bom desempenho, robustez e qualidade na entrega das sugestões. A ideia geral é criar um ciclo onde, a medida que os usuários interagem com o sistema, melhor ele vai ficando. Na Figura 3.1 mostramos a arquitetura geral do SCAC.

Tudo começa com o Gerente da Loja disponibilizando o catálogo de produtos que sua loja está vendendo no momento. Nessa hora, não temos consultas dos usuários, já que nenhuma interação com o sistema foi feita. A partir do catálogo, geramos consultas candidatas que passam por um filtro para obtermos uma lista de sugestões de consulta (mais detalhes sobre esse processo são mostrados na Seção 3.2). Com essa lista, conseguimos colocar em funcionamento o Serviço de Complementação Automática de Consultas (*Query Autocompletion Engine*), o qual usa a árvore de prefixos (ver detalhes na Seção 3.3) para armazenar e buscar sugestões em tempo real para os usuários que começam a digitar suas consultas na Caixa de Busca. Estas consultas digitadas também são capturadas e armazenadas para que sejam usadas no processamento offline, gerando consultas cada vez melhores e reiniciando-se o ciclo.

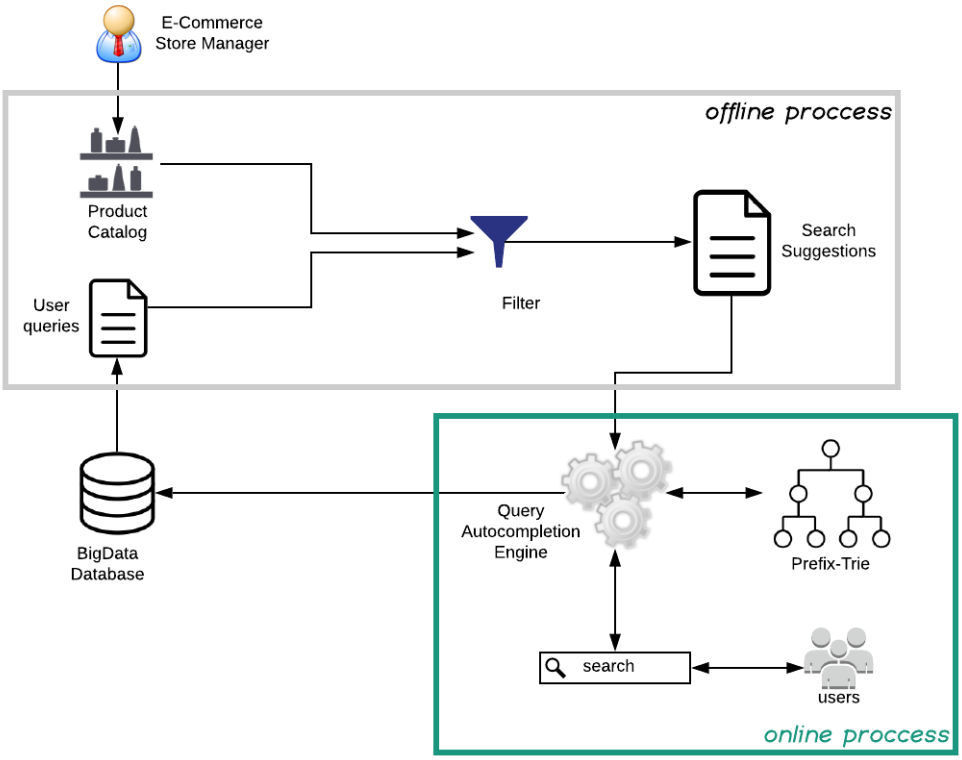


Figura 3.1. Arquitetura do Sistema de Complemento Automático de Consultas (SCAC) utilizado nos experimentos

É importante que haja esse ciclo porque um SCAC é um sistema dinâmico. Frequentemente produtos novos são inseridos no catálogo e produtos antigos saem de circulação. O mercado muda constantemente, assim como o desejo dos usuários pelos produtos. Consultas surgem e passam a ser muito buscadas em questões de instantes e

o sistema precisa se adaptar a esse ambiente. A seguir, mostramos com mais detalhes como é feito o processamento offline de geração de consultas.

3.2 Geração de Consultas

As consultas mostradas pelo SCAC para os usuários, enquanto estes digitam uma consulta, devem ser geradas previamente, no que chamamos de *processo offline*. Diz-se *offline* pelo fato dessas consultas já precisarem ter sido geradas previamente, antes do Serviço de Completação de Consultas, que responde às requisições por sugestões a partir do prefixo, esteja no ar. Esta geração envolve vários passos e segue algumas regras. Chamaremos esta etapa, daqui em diante, de Gerador de Consultas. O Gerador de Consultas é responsável por gerar as consultas que serão mostradas no processamento do prefixo a partir do catálogo de produtos e do comportamento dos usuários.

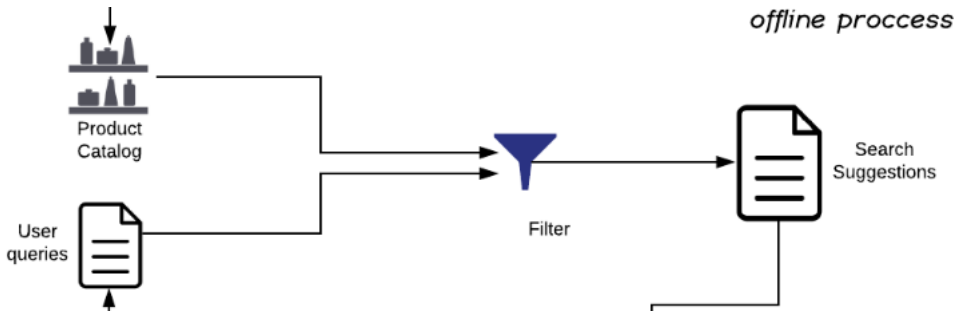


Figura 3.2. Gerador de Consultas: Primeiramente, o Gerador de Consultas obtém as informações das suas fontes e gera consultas candidatas. Na segunda etapa, as consultas candidatas passam por uma camada de filtro, onde são consideradas algumas regras para que a consulta seja validada.

Duas fontes para o Gerador de Consultas foram usadas nesta dissertação: o **catálogo de produtos** e as **buscas realizadas pelos usuários**. O catálogo de produtos é uma lista de documentos estruturados que possuem informações sobre todos os produtos que a loja em questão vende enquanto que as buscas realizadas pelos usuários são uma lista contendo tais consultas com suas respectivas frequencias dentro de um período especificado.

Mais adiante mostraremos dois experimentos relacionados com a escolha da fonte geradora de consultas. O primeiro (ver Seção 4.1) é um estudo das buscas que foram realizadas pelos usuários dentro de um período para descobrir em quais campos os termos das buscas estão. Por exemplo, os termos da busca *notebook samsung* podem ser encontrados nos títulos dos produtos (*notebook*) e no atributo marca (*samsung*). Outro experimento (ver Seção 4.3.1) foi realizado para entender, das consultas clicadas

no SCAC, quais foram geradas em cada fonte. Com essa resposta, conseguimos dizer qual é a melhor fonte geradora de consultas para o SCAC. A seguir, entraremos em mais detalhes sobre cada uma das fontes geradoras de consultas.

3.2.1 Catálogo de Produtos

Em lojas de comércio eletrônico, os alvos dos usuários são os produtos. Cada loja possui um catálogo de produtos referente aos produtos nela vendidos e os Sistemas de Busca e SCAC precisam auxiliar os usuários em sua tarefa de comprar o que desejam (e, às vezes, até o que nem sabiam que desejavam). Este catálogo é um componente ímpar de uma loja de comércio eletrônico e uma das principais diferenças em relação ao universo da WEB em geral. Diferentemente dos sites da WEB, o catálogo de produtos possui - quase sempre - uma estrutura bem definida que veremos a seguir. Além de ser uma fonte de consultas, é do catálogo de produtos, também, que conseguimos extrair informações úteis para a pontuação das consultas, como veremos na seção 3.4.

Dentre as informações do catálogo estão o título do produto; a sua descrição; o(s) seu(s) preço(s); a marca; a(s) hierarquia(s) de categoria(s); informações técnicas do produto que variam de produto a produto como tamanho da tela e quantidade de memória para celulares; quantidade de portas e capacidade para geladeiras, tipo de tecido e quantidade de lugares para sofás; a média das avaliações dos usuários, e etc. A tabela 3.1 mostra um exemplo real da estrutura de um produto no catálogo de uma loja. Muitas dessas informações são bastante úteis para gerar consultas para o SCAC. Entretanto, é preciso modificar e ajustar essas informações para estarem de acordo com o que as pessoas geralmente buscam. Um smartphone com o título *iPhone 6s Apple com Tela 4,7" HD com 128GB, 3D Touch, iOS 9, Sensor Touch ID*, por exemplo, não é buscado dessa mesma forma pelos usuários, mas estes utilizam termos mais gerais como *iphone* ou *iphone 6s* para encontrar tal produto. Aliás, esse mesmo produto, em uma loja online real foi clicado quando os usuários buscaram pelos seguintes termos que aparecem na Tabela 3.2

Desta forma, para gerar consultas a partir do catálogo, é preciso selecionar alguns campos. O algoritmo de geração de consultas a partir do catálogo desenvolvido neste trabalho foi feito de forma configurável para ser possível utilizá-lo diferentemente para cada segmento de loja, pois cada segmento tem um comportamento, atributos e características diferentes. Em lojas de moda, percebemos, analisando o comportamento dos usuários (mais detalhes na Seção 4.1) que a cor e categoria são atributos bem importantes para serem usados como insumo na geração de consultas, enquanto que em livrarias, o título, autor e editora são essenciais. Entretanto, alguns campos

Tabela 3.1. Exemplo de informações que um produto pode ter no catálogo

Campo	Valor
id	9040870
name	Notebook Dell Inspiron Ultrafino i15-7572-M20S 8ª Geração Intel Core i7 8GB 1TB Placa Vídeo 15.6' Windows 10
status	AVAILABLE
oldPrice	5799
price	5269
installment_count	9
installment_price	585.44
url	<url_link>
created	2018-03-14 10:32:34
image	<url_image>
eanCode	7899864914249
description	Mais real impossível. O novo Inspiron 15 7000 oferece um design elegante e Tela Infinita que amplia seus sentidos, mantendo a sofisticação e medidas compactas. Com acabamento em alumínio escovado, este produto possui uma qualidade incrível e oferece maior durabilidade, além do contorno lapidado cuidadosamente como diamante para expressar um estilo único.
details_peso	2 kg
details_marca	Dell
categories	Informática > Notebook

Tabela 3.2. Diferentes buscas que levaram uma quantidade de usuários a clicarem no produto *iPhone 6s Apple com Tela 4,7" HD com 128GB, 3D Touch, iOS 9, Sensor Touch ID* em 30 dias de log em uma loja online real.

Consulta	Quantidade de cliques
iphone 6s	765
iphone	476
iphone 6	431
iphone 6s 128gb	210
iphone 7	97
iphone 6s 128	54
iphone 6s 128 gb	24
6s	23
iphone 6s dourado	20
iphone 6s 64gb	12

são tão importantes que aparecem em praticamente todos os nichos: é o caso do título de produto e marca.

Outra técnica desenvolvida na geração de consultas a partir do catálogo de produtos, é utilizar técnicas de divisão em n-termos. Cada termo de um campo do catálogo e a combinação de N termos adjacentes pode gerar uma consulta diferente. Na Tabela 3.3 temos exemplos de consultas oriundas de divisões de n-termos, onde $n = 1$, $n = 2$ e $n = 3$. Estudamos o valor de n para cada nicho de loja, a fim de configurarmos de modo ideal a quantidade de termos adjacentes considerados para cada cenário (Mais detalhes do resultado deste estudo na Seção 4.1). A divisão mostrada no exemplo da Tabela 3.3 foi feita usando o nome do produto, mas pode ser feita para outros campos do catálogo, como categorias, atributos, marcas, autores, etc. Também geramos consultas combinando dois ou mais campos do produto. Combinar o nome com algum atributo, por exemplo, pode gerar consultas da forma *Harry Potter J. K. Rowling*, que é a junção de uma parte do nome com o autor. Outros exemplos são mostrados nas tabelas 3.3, 3.4 e 3.5.

Utilizar o catálogo de produtos para gerar sugestões de busca é importante pelo fato de que a sugestão sempre terá resultados, ou seja, como todas as palavras que utilizamos para gerar as consultas são indexadas, a consulta gerada, com certeza, terá, ao menos, um produto retornado pela busca (no mínimo, o próprio utilizado para gerar a sugestão), diferentemente das sugestões geradas pelo Comportamento dos Usuários que veremos a seguir, na Seção 3.2.2. Outro motivo de grande importância dessa fonte é que a obtenção desses dados é mais simples do que a extração do comportamento dos usuários. Geralmente o lojista já possui e fornece o catálogo de produtos

Tabela 3.3. Divisão do título de um produto em N-Termos. Livro Harry Potter Box Set Special Edition

Tamanho=1	Tamanho=2	Tamanho=3
Livro	Livro Harry	Livro Harry Potter
Harry	Harry Potter	Harry Potter Box
Potter	Potter Box	Potter Box Set
Box	Box Set	Box Set Special
Set	Set Special	Set Special Edition
Special	Special Edition	
Edition		

Tabela 3.4. Geração de consultas com mais de um campo do catálogo, usando combinação entre si. Um a um. Título: Notebook Sony Vaio. Atributos: 8gb, 16' e 500GB

Notebook 8gb	Sony 8gb	Vaio 8gb
Notebook 16'	Sony 16'	Vaio 16'
Notebook 500gb	Sony 500gb	Vaio 500gb

Tabela 3.5. Geração de consultas com mais de um campo do catálogo, usando combinação entre si. Dois por um. Título: Notebook Sony Vaio. Atributos: 8gb, 16' e 500GB

Notebook Sony 8gb	Sony Vaio 8gb
Notebook Sony 16'	Sony Vaio 16'
Notebook Sony 500gb	Sony Vaio 500gb

de forma estruturada, enquanto que o comportamento dos usuários ainda precisa ser extraído, armazenado e processado. Por conseguinte, estes últimos são mais difícieis de serem implementados pelos lojistas. Soma-se à importância dessa fonte, o fato dela ser independente do ciclo. Ao contrário das consultas oriundas do Comportamento dos Usuários que só existem a partir do momento que os usuários começam a buscar, as consultas geradas pelo catálogo de produtos existem sem essa dependência e são muito importantes para o processo de *cold start*, que é o nome que se dá ao momento de início do funcionamento do sistema na loja.

Gerar sugestões do catálogo com essas divisões e combinações de campos pode, muitas vezes, gerar consultas sem sentido como *box set*, mostrado na Tabela 3.4, por exemplo. Consultas *ruins* podem ser evitadas acrescentando regras ao filtro que falaremos mais adiante, na Seção 3.2.3.

3.2.2 Comportamento dos Usuários

Outra forma de gerar sugestões do Complemento Automático de Consultas é utilizando as próprias consultas dos usuários. O histórico de buscas dos usuários é uma mina de informações a ser explorada. Com ele conseguimos não só gerar *boas* consultas, mas também gerar pontuações sobre as mesmas com o uso da frequência com que essas consultas ocorrem em intervalos de tempos determinados. Diferentemente, as consultas geradas pelo catálogo de produtos não possuem informação intrínseca de relevância.

Aliás, não é só o histórico de buscas, mas toda a interação entre os clientes e a loja é uma fonte rica de informação. No varejo físico, algumas informações são extremamente difíceis de coletar, como por exemplo o número de visitantes da loja por hora / dia / semana / mês / ano, quais produtos cada cliente se interessarou em comprar, o que eles viram, por qual parte da loja cada cliente se interessa mais, dentre outras coisas. Algumas lojas, com alguma dificuldade, ainda conseguem mensurar dados mais genéricos como a quantidade de itens vendidos de um determinado produto e, às vezes, o histórico de compra de um cliente. No comércio eletrônico é possível extrair e armazenar todas essas informações citadas, aumentando assim extraordinariamente o conhecimento que a loja tem a respeito de seus clientes. Saber como capturar e manipular essa informação é um desafio extra para o lojista.

Extrair, guardar e explorar as buscas dos usuários não é uma tarefa trivial. É preciso ter um sistema robusto de captura e processamento de uma grande quantidade de dados em tempo real. Dependendo da loja, é possível em um só dia, por exemplo, atender cerca de 7 milhões de buscas e mais de 20 milhões de requisições para o SCAC.

Para gerar as consultas dos usuários, utilizamos o map-reduce que obtem todas as consultas que foram submetidas ao sistema nos últimos N dias. O resultado é uma lista de pares de consulta e a frequência de buscas de cada uma, como representado na Tabela 3.6. Essas consultas são submetidas ao processo de filtragem para então estarem prontas para serem usadas pelo SCAC.

Consultas geradas a partir das próprias consultas dos usuários são importantes porque os usuários repetem as consultas entre si, ou seja, a consulta que um usuário faz hoje é bem provável de ele ou outro usuário fazer amanhã. Fizemos um experimento sobre essa questão e mostramos na Seção 4.3.1. Além disso, como falamos anteriormente, as frequências podem ser usadas na fase de pontuação das consultas.

A seguir, mostraremos o funcionamento do processo de filtragem das consultas e sua importância para o SCAC.

Tabela 3.6. Lista de Consultas e Frequência em 90 dias de Log de uma loja real do nicho de Calçados Esportivos.

Consulta	Frequência em 90 dias
tenis nike	1075313
tenis adidas	770408
mochila	725037
tenis feminino	646190
tenis	477299
mochilas	364981
chuteira society	349274
nike	318997
tenis masculino	311789
tenis mizuno	274999
chuteira futsal	272237
tenis nike feminino	243917
bone	223686

3.2.3 Filtragem de Consultas

Todas as consultas candidatas geradas pela etapa anterior passam pelo processo de filtragem de consultas. Essa etapa é responsável por excluir *consultas ruins* e que não devem ser exibidas pelo SCAC para os usuários finais. O Sistema de Filtragem desenvolvido também é totalmente configurável para que se adeque às regras de negócio da loja em questão. A seguir detalharemos as regras desenvolvidas.

3.2.3.1 Conteúdo Impróprio

Algumas lojas possuem produtos de conteúdo impróprio e precisam que o Sistema de Complemento Automático de Consultas não sugira consultas e produtos deste segmento. São os casos de produtos de conteúdo adulto, sexshop, proibidos para menores de dezoito anos, etc. Também há casos de remédios que só podem ser comprados com comprovação médica, tarja preta, e que farmácias não podem sugeri-los. Criamos duas formas para filtrar essas consultas, a primeira é logo na geração de consultas baseadas no catálogo, onde excluimos consultas que foram geradas a partir de produtos categorizados como conteúdo impróprio; a segunda forma é submeter a consulta gerada para o Sistema de Busca e verificar se os produtos que retornam são de conteúdo impróprio. Para a segunda etapa, verificamos entre os N primeiros produtos se $X\%$ pertence à categoria imprópria. Os valores de N e X são configuráveis.

3.2.3.2 Início ou Término com StopWords

Existem consultas, geralmente geradas pelo catálogo de produtos, que começam ou terminam com *stopwords*[1]. A geração de consultas baseadas em n-termos durante o processamento do catálogo pode quebrar a frase antes ou depois de uma *stopwords*, fazendo com que a consulta gerada fique incorreta. São casos como "carrinho de", "de bebê", "livro da", etc.

Existem formas de descobrir se uma palavra é ou não uma *stopwords*, mas, neste trabalho, apenas configuramos uma lista de termos considerados *stopwords* manualmente. Na Tabela 3.7 vemos alguns termos considerados como stopwords neste trabalho.

Tabela 3.7. Lista de Consultas consideradas como exceções.

Termos
o
a
por
para
de
algum
the
to
for
el
la

Essa é uma regra que possui exceções. Existem consultas que começam ou terminam com *stopwords*, mas mesmo assim são consultas que não devem ser excluídas. Na Tabela 3.8 temos alguns exemplos das exceções.

Tabela 3.8. Lista de Termos considerados como StopWords nesta dissertação.

Consultas
the sims
o senhor dos aneis
el nino
the simpsons

[1]StopWords são palavras muito comuns que pertencem geralmente a classe de preposições e conjuções, como "de", "por", "para", etc.

Para reconhecer essas consultas, fizemos um validador que verifica se essa consulta foi realizada mais do que X vezes durante os N últimos dias e se essa consulta retorna mais do que Y produtos. Onde N, X e Y são valores configuráveis.

3.2.3.3 Composição apenas por caracteres numéricos

As consultas geradas pelo catálogo podem conter apenas caracteres numéricos. São casos onde está configurado para gerar consultas a partir de um atributo numérico como *eancode*, *isbn*, *id do produto*, etc. Donos das lojas e balconistas de lojas físicas, por muitas vezes, utilizam a loja virtual e o campo de busca para encontrar produtos que estão querendo visualizar. Como são *usuários diferenciais* que conhecem a loja, as buscas geralmente são feitas usando o *identificador do produto*, geralmente composto por apenas caracteres numéricos. O sistema de rastreamento e armazenamento coleta essa busca como se fosse uma outra qualquer e essas consultas numéricas acabam sendo geradas no processamento offline.

Para que essas consultas não apareçam no SCAC para o usuário final, desenvolvemos uma filtragem que valida se a consulta possui apenas caracteres numéricos e remove consultas que se encaixam nessa regra. Também validamos se a consulta pertence a algum atributo que seja um identificador de produto para também removê-las. Isso deve ser feito pois o usuário final não conhece esses códigos e não faz sentido mostrá-los.

Essa é uma outra regra que possui exceções. Existem consultas numéricas que são relevantes para os consumidores. É o caso de consultas como *2012*, *1808*, *007* que são filmes e livros que são buscados regurlamente pelos usuários de lojas que vendem tais itens. Essas consultas são identificadas pelo mesmo processo das *stopwords*, validando a frequência de busca dos últimos dias e a quantidade de produtos retornados.

3.2.3.4 Quantidade de Termos e Caracteres

Outra regra criada para validar consultas na etapa de filtragem é a quantidade de termos e caracteres que a consulta gerada possui. Consultas com muitos termos e/ou com muitos caracteres são excluídas da lista final. As quantidades limites de termos e caracteres são configuráveis.

3.2.3.5 Consultas Corretas

O Sistema de Busca desenvolvido possui quatro tipos de retornos para buscas textuais: (1) *Found*, (2) *Correction*, (3) *Approximate* e (4) *Not Found*. Buscas *Found* são consultas que possuem resultados exatos, em que o sistema não precisou interferir para

achar resultados. *Correction* são consultas onde o usuário buscou algum termo escrito de forma incorreta e o sistema precisou corrigir a consulta para encontrar resultados. *Approximate* são consultas onde, mesmo estando ortograficamente corretas, não possuem resultados exatos e, por causa disso, o sistema de busca decide retirar algum termo da consulta, aproximando-a para outra. Em último lugar, consultas *Not Found* são aquelas onde não há resultados e o sistema de busca não conseguiu nem corrigir nem aproximar.

Desenvolvemos também, na etapa de filtragem, regras para permitir que consultas sejam consideradas válidas apenas se forem encaixadas em um ou mais tipos de retorno de buscas, configurado manualmente. Geralmente, consideramos apenas buscas que são *Found*. Para isso, as consultas candidatas são submetidas ao sistema de busca, verificando se o retorno se encaixa em algum retorno válido configurado.

3.2.3.6 Consultas sinônimas

Como vimos na Seção 2.3, Cai & de Rijke [3] utiliza o método de conjuntos para identificar consultas sinônimas, verificando se uma consulta é similar a outra, apenas permutando os termos que a compõem. Neste trabalho, ampliamos o conceito de consultas sinônimas para entender consultas que são tão parecidas que diferem em número e grau, além da permutação de termos. Para o SCAC, consultas sinônimas são ruins para serem exibidas, já que elas levam a um mesmo resultado e acabam não contribuindo com nenhuma informação para a lista de sugestões. Na verdade, essas consultas são até prejudiciais para o sistema que perde espaços de consultas que poderia ter sido substituída por outras mais importantes. Assim, os pares *notebook asus* e *asus notebook*, *blusa vermelha* e *blusa vermelho* e *sofá* e *sofás* são consideradas consultas sinônimas.

Identificamos consultas sinônimas na etapa offline do processamento, ainda durante o processo de filtragem, mas, nesse caso, não excluimos consultas sinônimas, apenas identificamos, para cada consulta, quais são suas sinônimas. Durante o processamento do prefixo, antes de uma sugestão entrar na lista de respostas, verificamos se não há nenhuma consulta sinônima a ela já na lista e, só então, a adicionamos na resposta, garantindo, assim, que todas as sugestões da lista não sejam sinônimas entre si.

A decisão de não excluir consultas sinônimas se dá pelo fato de não sabemos qual é a versão principal de uma consulta. Na verdade, não podemos dizer que existe uma versão principal e as outras consultas são sinônimas, pois a consulta principal vai depender do prefixo digitado pelo usuário. Assim, se o usuários digitar *notebook as*, a

sugestão *notebook asus* deve aparecer e a sugestão *asus notebook* deve ser removida, mas se outro usuário digitasse *asus note*, o contrário deve acontecer.

3.3 Processamento do Prefixo

A etapa de Processamento do Prefixo ocorre quando um usuário começa a digitar algo na Caixa de Busca e o SCAC retorna sugestões que completam a consulta que o usuário acabou de digitar. O processo de buscar sugestões para o prefixo solicitado é o alvo desta etapa. Como vimos na Seção 2.2, Chaudhuri & Kaushik [5] desenvolveu uma técnica para obter sugestões de consulta já tolerantes a erro de digitação. O algoritmo base feito pelo autor para obter essas sugestões foi desenvolvido neste trabalho com poucas alterações.

Uma das alterações que propomos neste trabalho é quebrar as consultas em termos e inseri-los individualmente na árvore de prefixos, mas com a referência para a consulta completa. Essa técnica permite que, ao buscar por um prefixo, retorne sugestões em que o termo que completa o prefixo esteja em qualquer posição da consulta. Como veremos mais tarde, usaremos a posição da consulta e do prefixo para gerar pontuações para a consulta. O exemplo da Figura 3.3 mostra uma árvore que possui as consultas "samuel eto"e "notebook samsung"inseridas por termo. Se um usuário digitar o prefixo "sam"na caixa de busca, ambas as consultas serão retornadas, mesmo que uma delas ("notebook samsung") não tenha o primeiro termo iniciado por esse prefixo, mas o segundo. É importante salientar que o nó que contém a última letra precisa ter uma lista de ponteiros para as consultas verdadeiras.

Para que o Sistema de Complemento Automático de Consulta consiga sugerir consultas mesmo com erros de digitação, implementamos a busca na trie proposta por Chaudhuri & Kaushik [5] que permite encontrar sugestões com quantidade de erros limitada. Ao implementarmos esse algoritmo, percebemos que essa quantidade não pode ser fixa pra qualquer prefixo. Se considerarmos sugestões com distância de edição igual a dois, por exemplo, pra um prefixo curto, sugestões com essa distância de edição deixariam passar quase todas as opções de consultas registradas no sistema, enquanto que para prefixos longos, esse valor seria bastante restritivo. Percebemos, portanto, que o limite máximo de distância de edição que devemos permitir é proporcional ao tamanho do prefixo. Em nosso sistema, a quantidade de erros permitidos, obedece à equação:

Seja:

- $prefix$, o prefixo digitado;

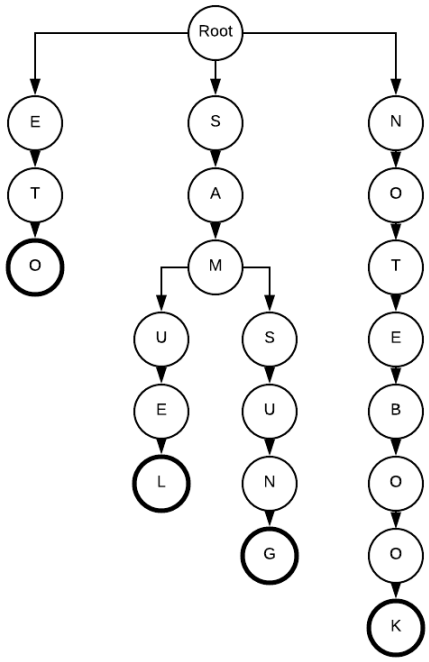


Figura 3.3. Exemplo de uma árvore de Prefixo com as consultas *Samuel Eto* e *Notebook Samsung* inseridas por termo.

- A função $len()$ representando o tamanho de caracteres da palavra, de modo que, por exemplo, $len('abcd') = 4$;
- MAX_ERROR uma constante configurável representando a quantidade máxima de erros permitido;
- $DIVISOR$ uma constante configurável, representando o valor que será dividido o tamanho do prefixo;
- $min()$ a função que retorna o menor valor entre dois números

O valor de erro máximo e permitido para um prefixo qualquer $prefix$ é dado por:

$$e = min(MAX\_ERROR, \frac{len(prefix)}{DIVISOR})$$

Controlamos a tolerância a erros que o sistema irá permitir pelas duas variáveis de configurações MAX_ERROR e $DIVISOR$. Quanto maior o MAX_ERROR, maior é o limite de erros permitidos para prefixos longos. Quanto maior o $DIVISOR$, menor é o erro permitido para o prefixo digitado. Na maioria das lojas, utilizamos a configuração:

- $MAX\_ERROR = 3$
- $DIVISOR = 4$

Na próxima Seção falaremos de todas as características que conseguimos extrair das consultas para gerar pontuações e como ordená-las para exibir as melhores consultas para os usuários.

3.4 Pontuações

Existem muitos fatores que influenciam o cálculo da importância de cada sugestão de consulta durante a fase de ordenação das sugestões para um prefixo. O Sistema de Complemento Automático de Consultas precisa não apenas saber quais sugestões completam o prefixo digitado, mas também determinar qual é a importância de cada sugestão e ordená-las para que as melhores sugestões apareçam sempre nos primeiros lugares.

O conceito de importância e a ordenação da lista de sugestões está sempre ligado a alguma métrica, ou seja, a ordenarção precisa ter algum alvo a ser atingido. Neste trabalho, as principais métricas que queremos aperfeiçoar são a taxa de uso do sistema e o MRR. Estas e outras métricas serão tratadas e explicadas na Seção 4.2 mais adiante.

A ordem das consultas sugeridas se faz ainda mais importante por causa da quantidade de espaços limitados de sugestões que temos para exibir. Esta quantidade de espaços varia de loja para loja, mas o padrão no e-commerce brasileiro são apenas 5 espaços disponíveis. O SCAC precisa usar esse espaço de maneira ótima.

3.4.1 Características de uma Consulta

Cada sugestão de consulta possui várias características e há uma pontuação relacionada com cada uma. Neste trabalho, as pontuações foram relacionadas com quatro grupos de características: o catálogo de produtos, o comportamento dos usuários, relacionadas com prefixo digitado e relacionadas com o usuário que digitou o prefixo. Deixamos o valor do peso de cada característica configurável para termos flexibilidade de testarmos várias combinações diferentes. A seguir, detalharemos como surgem e como calcular cada uma dessas características.

3.4.1.1 Características relacionadas com o Catálogo de Produtos

Características relacionadas com o catálogo de produtos são aquelas que dependem exclusivamente dos produtos que a loja vende. Elas são obtidas durante o processa-

mento offline, junto com a geração das consultas. Uma das características geradas por esse grupo é a **quantidade de produtos retornados pela sugestão**. Consultas que retornam poucos produtos e, principalmente as que não retornam nenhum produto podem ser despriorizadas de acordo com o valor dessa característica. Um detalhe é que nem sempre consultas que retornam muitos produtos são as melhores para estarem nas primeiras posições da ordenação final. Para obter esse valor, basta submeter as consultas geradas para o Sistema de Busca da loja e verificar a quantidade de produtos retornados.

Durante a geração de consultas a partir do catálogo de produtos, utilizamos diversos campos, como vimos na Seção 3.2.1. Nessa etapa, a lista de campos a serem processados é obtida por uma configuração. Assim, **o campo que origina a consulta** é uma característica a mais usada no processo de ordenação das sugestões. Geralmente, consultas que são originadas do nome, categoria e marca do produto são mais relevantes do que consultas originadas de outros campos.

Outra característica relacionada com o Catálogo de Produtos e com o Comportamento dos Usuários é a **importância do produto** para a loja. Como vimos na Seção 3.2.1, conseguimos gerar consultas a partir de produtos do catálogo. Um produto é considerado importante de várias maneiras diferentes. Mas neste trabalho, o valor do produto está relacionado com a quantidade de visualizações e vendas do produto considerando-se os últimos 60 dias de acordo com a fórmula a seguir, considerando $Puchases$ como vendas e $Views$ como visualizações.

$$V_{prod} = (2 * Purchases) + Views$$

Assim, consultas que são geradas de produtos muito vistos e comprados possuem um alto valor para essa característica. Vale ressaltar, também, que utilizamos o valor de 60 dias de log para gerar essa informação, mas outras quantidade de tempo também são válidas e podem ser testadas. Deixamos essa atividade para trabalhos futuros.

3.4.1.2 Características relacionadas com o Comportamento dos Usuários

Nesta dissertação também propomos características das consultas que são relacionadas ao comportamento dos usuários do site. Entende-se por comportamento dos usuários a agregação das ações que os usuários fazem, individualmente, durante um intervalo de tempo. Essas características são obtidas, também, durante o processo offline de geração de consultas.

A família de características mais comum dentre as relacionada com o Comportamento dos Usuários é a que representa **a quantidade de vezes que a consulta foi**

buscada nos últimos N dias. Falamos em família de características pelo fato do valor de N poder representar vários valores diferentes. Se queremos dar uma importância para consultas recentes, basta dar um peso maior para a frequência em 7, 4 e 2 dias, por exemplo. Se queremos consultas mais *robustas*, damos mais pesos para consultas com frequência alta nos últimos 30, 45 e 60 dias. Aliás, são com essas quantidades de dias considerados que Bar-Yossef & Kraus [1] cria o algoritmo MPC (*Most Popular Completion*) usado como *baseline* neste trabalho.

Da mesma forma temos uma família de características relacionadas com a quantidade de vezes que as buscas são feitas, também podemos gerar mais duas famílias de características: a **quantidade de vezes que a busca levou a um clique em produto** e a **quantidade de vezes que a busca levou a uma venda de um produto**. Os cenários para obter essas características são parecidos: Suponha que um usuário fez uma busca e visualizou vários produtos. Se ele clicar em qualquer produto do resultado, significa que aquela busca levou a um clique e então incrementamos o valor da quantidade de vezes que a busca levou a um clique em produto. De forma semelhante, se o usuário, além de clicar em algum produto, chegar a comprá-lo, incrementamos a quantidade de vezes que a busca levou a uma venda de um produto. Juntando com a quantidade de vezes que a consulta foi realizada, são três famílias importantes na hora de validar se a consulta é importante, representando o funil de compra dos usuários. Na Tabela 3.9 mostramos a quantidade dessas três métricas para algumas consultas em uma loja de comércio eletrônico real.

Tabela 3.9. Exemplos de Quantidade de Buscas realizadas, Quantidade de Buscas que levaram a um Clique e a Quantidade de Buscas que levaram a uma compra em uma loja de Comércio Eletrônico real durante 24 horas.

Consultas	Buscas	Cliques	Compras
geladeira	1249	469	1
microondas	709	332	7
fogao	482	192	0
retro	421	38	1
cooktop 5 bocas	281	158	5
lavadora	204	80	3
ative	195	43	1
pecas	177	35	7

Outra família de características pode ser gerada quando utilizamos a **taxa de clique** das consultas dos últimos N dias. A taxa de clique é uma característica que envolve a quantidade de vezes que uma busca é feita e a quantidade de vezes em que há clique em algum produto resultante dessa busca através da fórmula:

$$Tx_{click} = \frac{Qtd_{cliques}}{Qtd_{buscas}}$$

Essa característica, intuitivamente, traz um conceito melhor de importância por não só indicar a quantidade de vezes que uma busca é feita, mas por penalizar a busca se a mesma não levar os usuários que a fizeram a interagirem com os produtos mostrados. Na Seção 4.3.2 mostramos os resultados dos experimentos relacionados com a Taxa de Clique.

De maneira similar à taxa de clique, testamos mais duas famílias: a **Taxa de Venda por Clique** e **Taxa de Venda por Busca** variando a quantidade de dias de log. Essas duas taxas estão especificadas, respectivamente, nas fórmulas abaixo:

$$Tx_{v/c} = \frac{Qtd_{vendas}}{Qtd_{cliques}}$$

$$Tx_{v/b} = \frac{Qtd_{vendas}}{Qtd_{buscas}}$$

Assim como a Taxa de Clique vista anteriormente, essas duas características priorizam consultas que levam os usuários mais a fundo no funil de compra.

3.4.1.3 Características relacionadas com o Prefixo Digitado

Algumas características das consultas estão estritamente relacionadas com o prefixo digitado pelo usuário. Obviamente, essas características só podem ser calculadas quando o SCAC recebe o prefixo que precisa ser completado na etapa online. Utilizando o prefixo da requisição de algum usuário, o SCAC consegue gerar cinco características importantes e básicas que ajudam bastante na ordenação das sugestões. São características que gerarão pontuações baseadas no prefixo: **quantidade de correspondências**, **quantidade de correspondências nas posições corretas**, a **distância de edição**, a **quantidade de cliques na sugestão para este prefixo** e a **quantidade de vendas que a consulta levou para este prefixo**.

A quantidade de correspondências é o número de vezes que um termo do prefixo casou com um termo da sugestão, sem importar a posição. A quantidade máxima de correspondências que uma sugestão de consulta pode ter é a quantidade de termos que o prefixo possui. Isso só acontece quando a configuração da árvore de prefixos permite inserção e busca quebrando as consultas em termos, conforme vimos nas Seções 2.2 e 3.3. Se a árvore de prefixos não quebrar as consultas dessa forma, as sugestões só terão uma única correspondência. A obtenção dessa característica é feita quando quebramos o prefixo em termos e, para cada termo, buscamos sugestões na árvore

de prefixos. Todas as sugestões que completarem cada termo, têm sua quantidade de correspondências incrementada. Ao final do processamento de todos os termos do prefixo, teremos sugestões com diferentes quantidades de correspondências. Nas Tabelas 3.10 e 3.11 temos exemplos do cálculo dessa característica.

Outra característica das sugestões relacionada com o prefixo requisitado é a quantidade de correspondências nas posições corretas. Essa característica está bem relacionada com a característica anterior, a diferença de levar em conta a posição do termo do prefixo em relação à posição do termo da sugestão. Na Tabela 3.11, vemos que o prefixo "notebook as"gerou as sugestões "notebook asus"e "asus notebook". Essas duas sugestões têm quantidade de correspondências igual a 2, mas, para a sugestão "asus notebook", as correspondências estão nas posições erradas em relação ao prefixo. No prefixo, *notebook* é a primeira posição, enquanto nessa sugestão, *notebook* está na segunda posição, logo é uma posição incorreta. Sugestões com os termos nas mesmas posições em relação ao prefixo facilitam a visualização dos usuários e tal característica pode ser usada como um dos fatores determinantes na ordenação das sugestões do SCAC.

A distância de edição também é uma característica bastante importante para a ordenação das sugestões de busca no SCAC. Na Seção 2.2, vimos que uma melhoria da busca por sugestões na árvore de prefixos permite encontrar sugestões com erro de edição. Essa quantidade de erros está atrelada ao par prefixo-sugestão e pode ser usada como pontuação para compor também a ordenação das sugestões. Mostrar sugestões com erro de edição parece ser interessante apenas quando não há nenhuma sugestão 100% correta a se mostrar. Na Tabela 3.10, algumas sugestões possuem um erro de edição, é o caso de "samsung galaxy note"e "notepad", para o prefixo "noteb". Como visto, em casos onde há sugestões corretas para mostrar, consultas erradas não são boas opções. Mas podemos sugeri-las em casos onde não há nenhuma opção correta ou realmente não há o que o usuário está buscando, mantendo o usuário ativo no site.

Duas famílias de características relacionadas tanto com o prefixo digitado quanto com o comportamento dos usuário são a quantidade de vezes que uma busca foi clicada para um prefixo específico e a quantidade de vezes que uma busca levou a uma compra para um prefixo específico. Consideramos família de características porque podem estar relacionadas com N dias. Estas características são processadas em duas etapas. Na parte offline, geramos mapas do tipo o prefixo p levou a n cliques na consulta c e na parte online, obtemos o prefixo através da requisição e geramos a pontuação para essas características específicas.

As Tabelas 3.10 e 3.11 mostram dois exemplos das características relacionadas com os prefixos *noteb* e *notebook as*, respectivamente.

Tabela 3.10. Comparação das características relacionadas com o prefixo *noteb*.

Prefixo: *noteb*			
Sugestões	Qtd Correspondências	Qtd Posições Corretas	Distância de Edição
Notebook	1	1	0
Notebook Samsung	1	1	0
Notebook Vaio	1	1	0
Notebook Asus	1	1	0
Notebook 500GB	1	1	0
Notebook 14"	1	1	0
Capa para Notebook	1	0	0
Samsung Galaxy Note	1	0	1
Notepad	1	1	1

Tabela 3.11. Comparação das características relacionadas com o prefixo *notebook as*.

Prefixo: *notebook as*			
Sugestões	Qtd Correspondências	Qtd Posições Corretas	Distância de Edição
Notebook Asus	2	2	0
Notebook Asus 500GB	2	2	0
Notebook Asus 14"	2	2	0
Notebook Asus i7	2	2	0
Notebook 500GB Asus	2	1	0
Asus Notebook	2	0	0
Notebook Samsung	1	1	0
Notebook Vaio	1	1	0
Notebook	1	1	0
Notebook 500GB	1	1	0
Notebook 14"	1	1	0
Celular Asus	1	1	0
Samsung Galaxy Note	1	0	1
Notepad	1	1	1
Notepad Asus	2	2	1

3.4.1.4 Características relacionadas com o Usuário

Um outro campo que pode ser bastante explorado pelo SCAC é o Usuário. O SCAC deve ser capaz de entender quem é o usuário que está fazendo a consulta e propor sugestões de acordo com os seus gostos via sugestões personalizadas, a fim de aumentar seu engajamento e, consequentemente, gerar mais vendas para a loja. Há diversos estudos envolvendo personalização no SCAC, como vimos na Seção 2.3 e vamos explorar os conceitos de alguns deles neste tópico.

Antes de falar sobre o que pode ser obtido dos usuários, vale lembrar que fazer o SCAC funcionar de forma personalizada em sites de comércio eletrônico não é uma tarefa simples. Primeiramente, é preciso saber como capturar informações dos usuários e, depois, ter um sistema robusto de armazenamento e extremamente rápido para processar diversas informações e fornecer dados úteis em milissegundos.

Para este trabalho, procuramos apenas entender quais informações seriam úteis para personalizar as sugestões a nível de usuário, mas não implementamos a sua utilização, deixando a tarefa para trabalhos futuros.

Um tripé de características bastante úteis que podem ser extraídas dos usuários é o que propôs Shokouhi [11]: gênero, idade e localização. Em seu trabalho, Shokouhi [11], mostra que o MRR aumenta consideravelmente no SCAC para WEB em Geral e é bem provável que haja ganhos no comércio eletrônico também. Em lojas de moda, gênero e idade parecem ser muito úteis, enquanto que em lojas de departamento, a localização pode ser um diferencial.

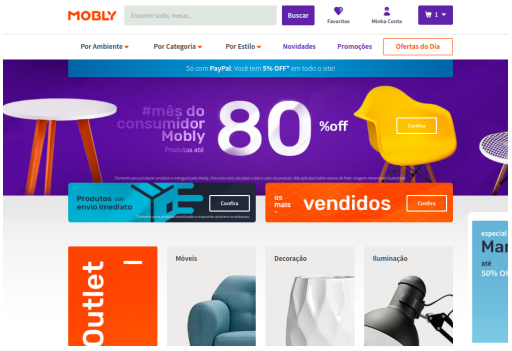
Um segundo nível de personalização também pode ser explorado quando agrupamos usuários que têm comportamento semelhante. Para isto acontecer, é preciso ter uma etapa offline de agrupamento para que, quando um usuário novo chegar no site, com poucos passos no site, consigamos descobrir a que grupo pertence e fornecer sugestões que pessoas de seu grupo receberiam.

Uma outra forma simples de personalizar o SCAC é saber quais consultas o usuário que está pesquisando no momento já fez anteriormente. Geralmente, usuários fazem a mesma consulta várias vezes. Então mostrar consultas que ele já fez ou parecidas com ela podem ter grande valia para o cliente final.

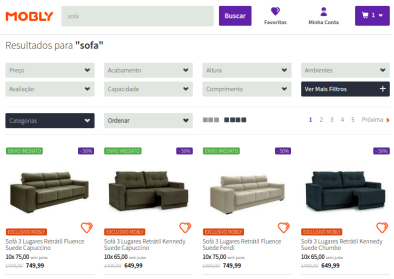
Da mesma forma que consultas anteriores, saber quais produtos o usuários em questão já viu ou comprou também é uma fonte importante para o SCAC. Tendo tal informação, é possível sugerir consultas que foram geradas por esses produtos ou sugerir consultas das categorias desses produtos.

É possível ainda utilizar a página em que o usuário está para melhorar as sugestões de consulta no SCAC. Na Figura 3.4, há várias páginas diferentes de onde é possível

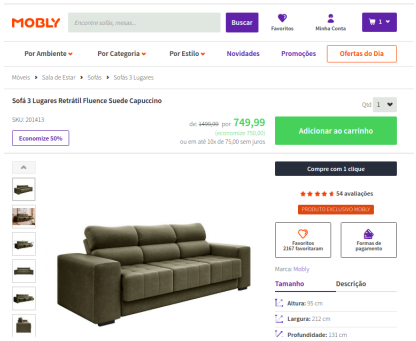
estar para realizar consultas. Com a página atual do usuário, pode-se dar prioridades a consultas específicas. Em uma página de categoria, por exemplo, consultas relacionadas com aquela categoria podem ter prioridade na ordenação. Se o cliente está na página da categoria referente a câmeras fotográficas, ao digitar "ca"no SCAC, pode-se dar preferências às consultas sobre "câmera"do que "cafeteiras"ou "camisa". Já numa página de produto, consultas relacionadas com aquele produto podem ganhar uma maior relevância na ordenação final.



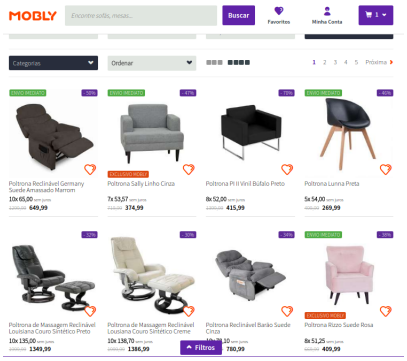
(a) Home



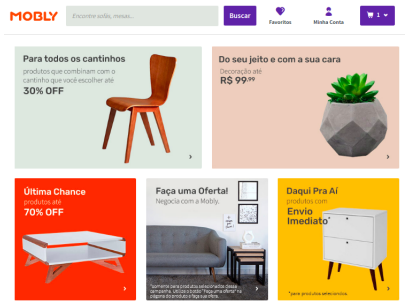
(b) Busca



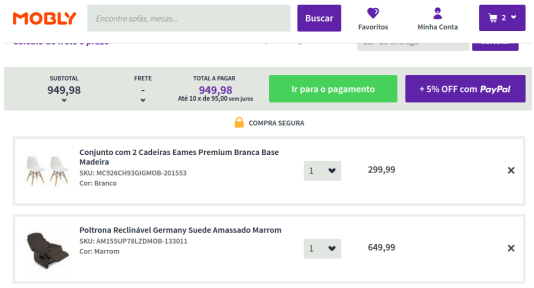
(c) Produto



(d) Categoria



(e) Hotsite



(f) Carrinho

Figura 3.4. Comparação de tipos de páginas de um E-Commerce (a) Home, página principal do site; (b) Página de Busca; (c) Página de Produto (d) Página de Departamento/Categoria; (e) Página de Hotsite; e (f) Página de Carrinho.

Em resumo, neste trabalho as seguintes características foram observadas, extraídas e calculadas para que se pudesse utilizar nos métodos de ordenação de consultas. Estão expostas na tabela 3.12

Tabela 3.12. Comparação das características extraídas

Catálogo	*Usuários*	*Prefixo*
QTD de Produtos	Frequencia	QTD de Correspondências
Campo	Cliques	QTD de Correspondências Posições Corretas
-	Vendas	Distância de Edição

Como foi visto, várias características podem ser usadas para melhorar o resultado da ordenação do SCAC. As que foram citadas são as mais simples de serem obtidas e as que temos conhecimento até o momento. Muitas outras podem ser geradas, extraídas e usadas para melhorar ainda mais a ordenação das sugestões. Saber como combiná-las se torna um desafio tão grande quanto extraí-las, ou seja, saber o valor e importância de cada uma delas não é uma tarefa trivial. Veremos alguns estratégias sobre como ordenar essas sugestões a seguir.

3.4.2 Manipulação das Sugestões pelo Lojista

Uma grande diferença do contexto de comércio eletrônico para Web é que o lojista pode querer influenciar na ordenação de sugestões. Para isso, ferramentas são criadas para que o cliente consiga configurar destaques como: para o prefixo p, a consulta *query* deve aparecer em primeiro lugar. Muitas vezes isso é necessário por conta de contratos dos lojistas com as marcas e o SCAC deve saber lidar com tais configurações. Em nossos testes, retiramos destaques desse tipo para não ter influência dos mesmos nos resultados.

Além de destaques em sugestões, o lojista também pode configurar que duas ou mais consultas sejam sinônimas. Assim, além das regras automáticas que são geradas no processamento offline, adicionamos as regras configuradas manualmente pelo lojista. Da mesma forma que os detaques, retiramos essas configurações nos testes executados. Outras duas ferramentas que o lojista consegue influenciar no SCAC é adição e remoção de consultas.

3.4.3 Ordenação das Consultas

Após verificar alguns exemplos de características que podem ser extraídas para servir de pontuação para cada sugestão de busca no SCAC, precisa-se saber como combiná-

las para gerar a pontuação final e, depois, ordená-las. A esse processo demos o nome de Ordenação das Consultas.

Existem várias formas de gerar ordenações para um conjunto de objetos em geral. Na verdade, a dificuldade não é na tarefa de ordenação em si, mas em como saber qual sugestão é a melhor baseando-se nas características de cada consulta. Vimos na Seção 2.3 algumas formas existentes para ordenar um conjunto de objetos usando Machine Learning. Neste momento, qualquer técnica de ordenação é válida e pode ser estudada e testada. De maneira geral, problemas de ordenação se resumem a ter uma lista de candidatos, onde cada um deles tem uma lista de características e o objetivo é saber qual é a melhor ordem, utilizando alguma métrica.

A métrica é o ponto chave em uma ordenação. No caso de lojas de comérécio eletrônico muitas métricas podem ser estudadas e focadas em métodos de ordenação. As métricas comuns para SCAC são o MRR (Mean Reciprocal Rank), e R@N (Revocação a N) para diferentes tamanhos de prefixos. Quando falamos de comércio eletrônico, outras métricas aparecem como RPV (Real por Venda), Conversão, Faturamento, dentre outras. Dependendo de qual métrica queremos melhorar, a ordenação dos candidatos pode ser diferente, focado nessa métrica.

Um dos métodos mais simples de ordenação é quando se olha para as características e escolhe-se um peso para cada uma delas. Assim, ao obter a lista de objetos, a ordenação final é calculada a partir do score final de cada objeto. Esse score é gerado multiplicando-se o peso de cada característica com o valor dela em cada sugestão. Esse método é chamado de Combinação Linear.

No caso da combinação linear, o desafio é encontrar quais são os valores ideais para os pesos em cada característica. Para isso, muitas técnicas são encontradas na literatura, mas que dependem de uma métrica referencial. Para conseguir chegar na melhor combinação dos pesos, é necessário utilizar os conceitos de Machine Learning. Dado uma classificação ideal baseada em alguma métrica (por exemplo, cliques em sugestões no SCAC), algoritmos de Machine Learning utilizam uma base de treino para encontrar a combinação ideal de pesos.

Neste trabalho, utilizamos a técnica de ordenação em duas etapas: durante o processamento offline, fazemos uma combinação linear das features que temos acesso neste cenário como: frequencia da consulta em N dias, quantidade de produtos retornados, influência do produto e do campo para a sugestão, dentre outras citados nas Seções 3.4.1.1 e 3.4.1.2. O resultado dessa combinação linear é um valor de importância das sugestões sem contexto, ou seja, sem depender de qual prefixo está sendo processado. Na etapa online, quando temos acesso a features relacionadas ao prefixo citados na Seção 3.4.1.3, usamos a árvore de decisão simples para ordenar as sugestões

candidatas. A árvore de decisão simples prioriza as características em uma sequência e, caso duas sugestões empatem em uma característica, o desempate acontece na próxima característica da sequência. A sequência usada foi:

- Maior Quantidade de Correspondências;
- Menor Distância de Edição;
- Maior Quantidade de Correspondências nas Posições Corretas;
- Maior Valor da Combinação Linear da Primeira Etapa.

A árvore de decisão simples é bem utilizada neste caso porque as três primeiras características possuem valores muito pequenos, próximos e que a diferença entre esses valores tem uma importância muito grande no SCAC. *Notebook*, por exemplo, apesar de ter um valor na primeira etapa bastante alto por ser muito bem requisitado e estar em bons produtos, não é mais importante que *notebook asus* para o prefixo *notebook as*. Outro exemplo, como já vimos em seções anteriores, a sugestão *asus notebook* ficaria estranha se viesse antes de *notebook asus* para o mesmo prefixo. Além do mais, o objetivo do SCAC é completar consultas, então, para o prefixo *notebook as*, consultas que começam com notebook e tem o segundo termo algo que complete o prefixo *as* são as melhores a serem exibidas e, provavelmente, é o que o usuário que está digitando deseja. No final das contas, o prefixo é tão importante que todas as features que se relacionam com ele, para esse método, são mais importantes que as demais.

Existem ainda outros algoritmos de ordenação que podem ser utilizados no SCAC. Árvores de decisão mais complexas, Deep Learning, Redes Neurais, SVM, entre outros são utilizados na literatura em diversos cenários. Todos já foram bastante estudados e tem prós e contras, dependendo do lugar onde são aplicados. Deixamos para realizar experimentos com essas outras técnicas de ordenação em trabalhos futuros.

Capítulo 4

Experimentos e Resultados

Este capítulo tem por finalidade mostrar todos detalhes dos experimentos realizados e dos resultados obtidos. Está dividido em três Seções. Em 4.1 mostramos as configurações do ambiente em que os experimentos foram realizados, assim como os segmentos das lojas que foram submetidas à experimentação. Na Seção 4.2 falaremos das métricas importantes para os experimentos e como calculamos cada uma delas. Em 4.3 mostramos os resultados dos experimentos e traçamos comentários sobre os mesmos.

4.1 Ambiente de Experimentação

Podemos dividir o ambiente dos experimentos em dois: ambiente de experimentação online e offline. O ambiente de experimentação online é aquele onde testamos dois ou mais cenários em uma loja de comércio eletrônico real, onde usuários reais fazem consultas ao sistema que tem capacidade de sortear o tráfego que chega e direcionar para qual algoritmo deve responder essa requisição, chamamos também de Teste A/B. O ambiente de experimentação offline acontece quando guardamos registros de prefixos e consultas submetidos ao sistema de buscas para avaliação posterior. Neste caso, consultas são realizadas em um ambiente controlado e não no cenário de produção.

Para os testes que foram executados em produção, utilizamos quatro máquinas com configurações iguais e um balanceador de carga. As quatro máquinas possuem 4 cores, Sistema de Arquivos em SSD de 500GB e 30GB de RAM. Já os testes offline foram executados em uma única máquina, 4 cores, processador Intel Core I5, 2,5GHz, 12GB de RAM e Sistema de Arquivos HD com 500GB.

O SCAC foi desenvolvido usando a linguagem C++, com bibliotecas padrões e originais da linguagem como STD, boost, dentre outras.

4.1.1 Caracterização da Base de Dados de Teste

Para a realização dos experimentos, montamos uma base de dados de testes com dados extraídos de um conjunto de seis lojas reais de comércio eletrônico. As lojas utilizadas são lojas de diferentes segmentos de mercado, a saber: Lojas departamentais, Moda, Móveis, Cosméticos, Livrarias e Eletrodomésticos. As principais características de cada loja estão listadas na Tabela 4.1. Para cada produto de cada loja, a coleção de dados tem como campos estruturados informação sobre título, descrição e atributos estruturados que variam de produto para produto, tais como dimensões, cor, peso, dentre outros que foi visto na Seção 3.2.1.

A Tabela 4.1 apresenta dados gerais sobre a base de dados de produtos de cada uma das seis lojas utilizadas nos experimentos. Pode-se ver que as seis lojas representam um conjunto bem diverso quando se olha informação sobre suas respectivas bases de dados. Essa variação é importante para os experimentos. Por exemplo, lojas com grandes quantidades de produtos tendem a ter produtos menos conhecidos pelos usuários e geralmente são lojas onde catálogo de produtos muda frequentemente. A Quantidade de Acessos refere-se à quantidade de usuários que visitam a loja, o valor na tabela refere-se à quantidade média de buscas em um dia comum. A Diversidade de Produtos faz referência a quantos produtos de categorias diferentes a loja vende. A loja de eletrodomésticos, por exemplo, vende poucas categorias diferentes, enquanto que a loja de departamentos naturalmente vende uma variedade muito maior de produtos, indo desde relógio, celular, notebook, até a roupas, brinquedos, e comésticos. É importante ressaltar que nos experimentos online realizados as quantidades podem variar ligeiramente de um dia para o outro em função de mudanças no estoque da loja. Os valores apresentados servem, entretanto, de referência para entendermos características importantes de cada uma das lojas apresentadas nos experimentos.

Pode-se notar, por exemplo, que temos na coleção uma boa variedade tanto no tamanho da base de dados, quanto no número de acessos e na diversidade de produtos representada pelo número de categorias de produtos disponíveis.

Por exemplo, a loja do segmento de eletrodomésticos tem apenas cerca de mil produtos e apenas 5 categorias, enquanto tem uma quantidade média de acessos de 25 mil acessos diários. Já a do segmento de livraria tem uma base de produtos muito grande com pouca diversidade de produtos e cerca de 135 mil acessos. A departamento tem muita diversidade de produtos e grande quantidade de acessos. A loja de cosméticos tem pouca variedade e quantidade de produtos, mas um tráfego razoável. A loja de moda tem uma quantidade maior de produtos e acessos semelhante do que a loja de cosméticos, enquanto que a loja de móveis tem uma quantidade de produtos e acessos

maior. Ainda que outros cenários possam ocorrer em lojas reais, com as diferenças apresentadas poderemos estudar o desempenho de nossas propostas em diversos cenários distintos e observar como as variações nos cenários podem afetar o desempenho dos métodos estudados.

Tabela 4.1. Informação gerais sobre a base de dados de produtos de cada uma das seis lojas utilizadas nos experimentos: tamanho do catálogo (número de produtos), quantidade de acessos, número de categorias e nicho de mercado.

Segmento	Tamanho do Catálogo	Quantidade de Acessos	Diversidade de Produtos
Eletrodoméstico	1k	25k	5
Livraria	2,8M	135k	15
Departamento	1,0M	2M	+400
Cosméticos	4,8k	215k	+20
Moda	39k	190k	+20
Móveis	138k	290k	+40

Além das informações a respeito da base de dados de produtos, a coleção de experimentos criada contém também dados a respeito das consultas submetidas às lojas. As consultas foram extraídas de logs contendo consultas reais realizadas por usuários. A seguir são apresentados dados a respeito dos logs de acesso utilizados nos experimentos. Uma primeira informação importante diz respeito à ocorrência de termos das consultas nos diversos campos de informação que temos na base de dados sobre produtos.

A Tabela 4.2 apresenta informação sobre a ocorrência dos termos das consultas submetidas por usuários nos diversos campos de informação a respeito de produtos existentes em cada uma das lojas. Para o cômputo dessas estatisticas foram utilizados logs de 90 dias de consultas de cada loja e, para cada termo das consultas, verficou-se em que campos de produtos da base de dados da loja o mesmo ocorre.

Pela Tabela 4.2, podemos perceber que a Loja de Eletrodomésticos possui muitas consultas que não se enquadram no catálogo, atingindo porcentagem maior que as obtidas nas lojas dos demais segmentos: 8,8%. A maior concentração das buscas para essa loja está nos atributos dos produtos, com cerca de 85% dos termos buscados. É importante notar que a soma das porcentagens dos campos do catálogo não precisa ser 100%, já que um termo pode pertencer a mais de um campo.

As Lojas dos segmentos de Livraria e Departamento apresentam características similares. Em ambas, os termos são mais vistos nos títulos e nas descrições dos produtos

Tabela 4.2. Localização dos Termos buscados pelos usuários no Catálogo

Loja	Título	Descrição	Atributos	Nenhum
Eletrodoméstico	77%	82%	85%	8,8%
Livraria	92%	95%	73%	2,2%
Departamento	88%	88%	77%	3%
Cosméticos	88%	94%	97%	1%
Moda	87%	93%	93%	2%
Móveis	98%	97%	98%	0,5%

do que nos atributos. A taxa de termos que não estão em nenhum campo também é semelhante. A Loja do segmento de Cosméticos, tem a maioria dos termos buscados nos atributos de seus produtos. Possui também uma taxa de termos não encontrados bastante baixa, 1%, mostrando que é um catálogo bem preenchido. A Loja do segmento de Móveis e Decoração é a que tem o melhor catálogo, aparentemente. Os termos buscados estão presentes em quase todos os campos principais e a taxa de termos inexistente no catálogo é quase nula. É claro que o segmento pode influenciar, mas um catálogo bem feito com os campos claros e objetivos facilita a busca por produtos por parte dos usuários da loja.

A Tabela 4.3 apresenta as quantidades mínima, máxima e média de termos das consultas buscadas pelos usuários nas 6 lojas. Essa informação é útil porque o tamanho das consultas pode afetar diretamente decisões de projeto e o desempenho do SCAC. Por exemplo, se as consultas possuem muitos termos, pode ser interessante que o sistema priorize sugestões de consultas com mais termos.

Tabela 4.3. Quantidade mínima, máxima e média de Termos por Loja

Loja	Qtd Mínima	Qtd Máxima	Qtd Média
Eletrodoméstico	1	8	2,05
Livraria	1	10	2,52
Departamento	1	7	1,75
Cosméticos	1	6	1,64
Moda	1	6	2,22
Móveis	1	10	1,86

Vimos que as Lojas dos segmento de Departamentos e Móveis são as que recebem consultas com menor número médio de termos. A média mais alta aparece na loja do

nicho de Livraria, onde a maioria das consultas possui mais de um termo. Nas Lojas dos segmentos de Livraria e Móveis, respectivamente, a quantidade máxima de termos obtidos foi dez termos entre as três mil consultas mais frequentes. Em todas, o mínimo de termos buscado foi de um termo.

Outra forma de medir o tamanho das consultas é contar a quantidade de caracteres nas consultas ao invés do número de termos. A Tabela 4.4 apresenta os valores obtidos para cada loja. Apenas como observação de implementação, espaços em brancos que separam, e caracteres não alfa-numéricos como símbolos, hífen e outros também foram considerados como quantidade de letras digitadas. Letras acentuadas, no entanto, contam como apenas uma letra.

Tabela 4.4. Quantidade mínima, máxima e média de Letras (caracteres) por Loja

Loja	Qtd Mínima	Qtd Máxima	Qtd Média
Eletrodoméstico	1	50	12,86
Livraria	1	50	15,69
Departamento	1	48	11,15
Cosméticos	1	45	10,51
Moda	1	39	15,31
Móveis	1	50	11,88

Assim como a maior média de quantidade de termos, a Loja do segmento de Livraria também tem a maior média de letras digitadas dentre todas as lojas testadas, com média de 15,69 letras digitadas por consulta. A Loja de Cosméticos foi a que teve a menor média de letras digitadas, com 10,51 letras em média.

Um último dado importante que ajuda a caracterizar a coleção utilizada nos experimentos é a informação sobre como e se os usuários dessas lojas repetem uma mesma consulta mais de uma vez durante o processo de compra. Essa informação pode ser importante, por exemplo, para saber se a personalização do SCAC traria um ganho para o lojista, já que, na personalização das sugestões, consultas são mostradas de acordo com o interesse de um usuário, especificamente. Apesar do resultado, deixamos os experimentos de personalização do SCAC como trabalhos futuros.

Para essa análise, verificamos quantos usuários, em média, repetem uma mesma consulta durante uma sessão. O conceito de sessão é bastante difundido no E-Commerce, apesar de existirem diferenças entre os lojistas. Para o Google, uma sessão consiste em eventos contínuos de um usuário em que o tempo entre um evento e outro não pode passar de trinta minutos.

O resultado deste experimento está disponibilizado na Tabela 4.5. Surpreendemente, a loja do segmento de Moda obteve o maior índice de usuários que repetem a mesma consulta: cerca de 59%. Em segundo lugar, a Loja 3, de Departamentos obteve 56%. A loja de eletrodomésticos foi a que teve o menor índice de usuários que repetem a consulta, com 31%. Fica claro que utilizar essa informação é bastante importante para o SCAC, impactando quase sempre um terço dos usuários e, as vezes mais da metade deles.

Tabela 4.5. Porcentagem de usuários que repetem uma mesma consulta

Loja	Porcentagem
Eletrodoméstico	31%
Livraria	37%
Departamento	56%
Cosméticos	46%
Moda	59%
Móveis	48%

4.2 Métricas

Nesta seção mostraremos as métricas utilizadas nos experimentos. Essas métricas são importantes na hora de avaliar os resultados obtidos e comparar as diversas estratégias de ordenação para o SCAC.

4.2.1 Mean Reciprocal Rank

O MRR, do inglês *Mean Reciprocal Rank*, de um sistema avalia a classificação média recíproca dos cliques em sugestões. As sugestões são mostradas de acordo com uma ordenação, como visto nos capítulos anteriores. O MRR é uma média do inverso da posição da sugestão clicada, ou seja, quanto mais o sistema levar os usuários a clicarem nas primeiras sugestões, melhor é o sistema. Se os usuários clicam somente nas sugestões mais abaixo no ranking ou não clicam, o MRR deste sistema é baixo. Abaixo temos a sua fórmula.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i}$$

O cálculo é feito para cada evento. O RR (*Reciprocal Rank*) é calculado dividindo-se o número um pela posição do clique. A média entre todos os eventos é o MRR. Se um usuário clica na primeira sugestão mostrada, o RR vale 1. Se clica na segunda sugestão, o RR vale $\frac{1}{2}$. E assim por diante. No caso de um usuário não clicar em nenhuma sugestão, o RR é considerado 0 (zero). Para os experimentos realizados, cada letra digitada pelo usuários, fazendo uma requisição para o SCAC, conta como uma requisição no cálculo do MRR, ou seja, se o usuário não clica na sugestão, é computado o valor 0 (zero), deixando o valor de MRR bem baixo. O importante é comparar o MRR de cada técnica experimentada.

4.2.2 Média de Caracteres até o Clique

A média de caracteres diz respeito à quantidade de caracteres (letras, números, espaços, etc) que o usuário digitou até clicar em alguma sugestão de consulta. Com essa métrica, conseguimos saber se estamos mostrando as melhores sugestões de busca logo no início da digitação da consulta pelo usuário, ou seja, quanto menor a média de caracteres, mais rápido mostramos a sugestão que o usuário quer buscar. As técnicas de ordenação podem influenciar essa média.

essa métrica não pode ser utilizada de forma isolada. Como ela mede a média de caracteres digitados apenas no clique, uma média menor não significa que a técnica é melhor, ou seja, é interessante sempre acompanhar a taxa de clique em consultas no SCAC junto com essa métrica.

4.2.3 Média de Termos até o Clique

A média de termos mostra a quantidade de palavras digitadas pelo usuário até que ele clique em alguma sugestão. A média de termos de um sistema é útil para verificar se os usuários da loja online tendem a digitar muitas palavras antes de clicar em alguma sugestão.

4.2.4 Porcentagem de Cliques em Sugestão

A Porcentagem de Cliques em Sugestões (doravante no texto mencionada como PCS) do SCAC diz respeito à quantidade de cliques em relação à quantidade de buscas totais realizadas quando houve sugestões do SCAC. Uma busca pode ser realizada por vários meios dentro de um site, ou até de acesso externo. Para essa métrica, só contabilizamos o total as buscas realizadas quando o SCAC teve a chance de mostrar as sugestões.

Outro detalhe importante dessa métrica é que agrupamos essas buscas em pequenas sessões. Se não há uma interação com o sistema nos últimos 20 segundos, a sessão termina e a métrica começa a contar uma nova busca. Este procedimento é útil para que um mesmo usuário possa buscar várias vezes sem interferir no resultado da métrica.

$$PCS = \frac{Total_{cliques}}{Total_{interacoes*}}$$

A contagem é feita da seguinte forma: o número total de grupos de interações com o SCAC é o denominador desta fração que é incrementado toda vez que um usuário começa a digitar uma consulta para o SCAC. Se esse usuário parar de digitar por 20 segundos, e só depois voltar a digitar, o número total é incrementado novamente, mas se o usuário continuar digitando, com o intervalo entre cada caracter menor que 20 segundos, significa que este ainda não acabou de digitar e essas requisições são agrupadas, não incrementando o denominador em todas as requisições, mas apenas na primeira vez. O numerador dessa fração é a quantidade de cliques em sugestões do sistema. Com essa métrica, conseguimos saber do total de vezes que o SCAC é acionado, quantas ele leva o usuário a clicar em uma de suas sugestões. Quanto maior a porcentagem, melhor é o sistema.

4.2.5 Porcentagem de Usuários que Clicam em Sugestão

A Porcentagem de Usuários que Clicam em alguma Sugestão (de agora em diante referenciamos essa métrica por PUCS) do SCAC é uma métrica semelhante à PCS, diferenciando apenas na sessão utilizada. Enquanto na PCS a sessão é uma falta de interação do usuário por 20 segundos com o sistema, na PUCS a sessão é cada usuário. No final das contas, a métrica informa, do total de usuários que interagiram com o SCAC quantos usuários chegaram a clicar em alguma sugestão sugerida. A fórmula do cálculo dessa sugestão é dada por:

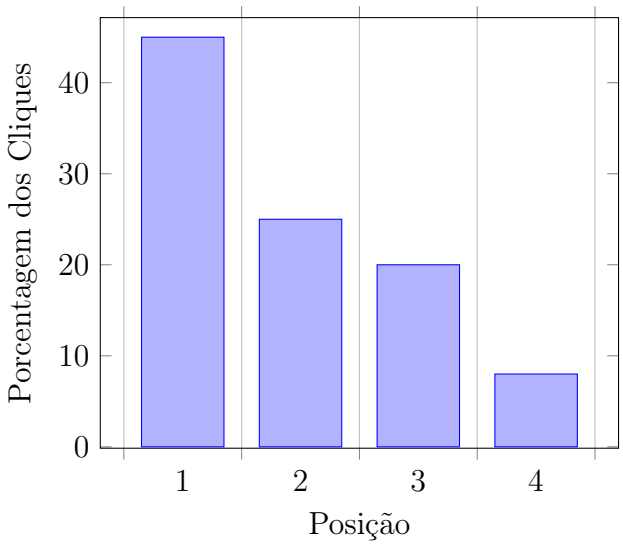
$$PUCS = \frac{Usuarios_{cliques}}{Usuarios_{interacoes*}}$$

4.2.6 Porcentagem de Cliques por Posição

Com o objetivo de medir a ordenação das sugestões do SCAC, mediremos, também, a Porcentagem de Cliques Por Posição (chamaremos de PCPP daqui em diante). Essa métrica nos diz como estão distribuídos os cliques em cada posição de ordenação, ou seja, do total de cliques em qualquer sugestão, essa métrica nos mostra qual a

porcentagem aconteceu quando a sugestão estava na primeira posição, na segunda, e assim por diante. Um bom algoritmo, de acordo com essa métrica, é aquele em que as maiores distribuições de cliques acontecem nas primeiras sugestões. Na Figura 4.1 mostramos um exemplo da PCPP real.

Figura 4.1. Gráfico com porcentagens de cliques por posição de Exemplo



4.2.7 Revocação

Outra métrica utilizada nos experimentos é a Revocação. A revocação, geralmente, está atrelada a uma posição limite n, representada pelo separador @. $R@5$, por exemplo, significa uma revocação até a posição 5. A revocação utilizada neste trabalho é um valor médio da fração de vezes que a consulta clicada está em uma das posições permitidas pelo total de requisições feitas ao sistema. Seria como se só existisse uma única consulta correta para o usuário, que é exatamente a consulta que ele irá fazer a busca. No SCAC ainda incluímos o tamanho do prefixo na revocação. Como um exemplo, suponha que um usuário digitou o prefixo *not* e deseja buscar por *notebook*, se o SCAC sugerir essa consulta entre as primeiras n posições, a revocação pra este prefixo é 1, senão é 0. Esse cálculo é feito para todas as consultas que serão testadas, variando os prefixos em todos os tamanhos possíveis.

Geralmente calcula-se essa métrica pra vários prefixos, tirando assim, uma média ao final que representa a Revocação média do SCAC até o limite n e para prefixos de tamanhos x.

4.3 Resultados

Nesta seção mostraremos os experimentos realizados e os resultados obtidos para as diversas estratégias de ordenação para SCAC em E-Commerce. Os experimentos foram direcionados através de uma comparação da fonte geradora de dados: Catálogo de produtos e Comportamento dos Usuários e também de uma exploração da fonte do Comportamento dos Usuários.

4.3.1 Fonte geradora de sugestões do SCAC: Catálogo de Produtos x Comportamento dos Usuários

Na Seção 3.2 vimos duas fontes geradoras de consultas para SCAC em Comércio Eletrônico: as buscas realizadas anteriormente pelos usuários e o catálogo de produtos. Também vimos os prós e contras em relação a como gerar essas consultas para cada fonte. Para as buscas realizadas anteriormente pelos usuários, a vantagem é que a informação já vem pronta e não precisa ser muito processada, mas, em contrapartida, é uma fonte de difícil extração e o lojista precisa ter um sistema robusto para processar todos os dados. Consultas geradas do catálogo de produtos têm a vantagem de serem mais fáceis de serem extraidas, já que o catálogo geralmente é fornecido pelo lojista. Uma desvantagem entretanto é a necessidade de combinar e criar diversos filtros para não gerar consultas ruins.

Para verificar qual é a melhor fonte geradora de consultas, fizemos o seguinte experimento: geramos as consultas a partir das duas fontes e marcamos todas as consultas dizendo se ela veio do catálogo de produtos e/ou de buscas anteriores de usuários. Note que neste momento, uma mesma consulta pode estar marcada que veio das duas fontes. Além disso, deu-se importância equivalente às duas fontes nesse experimento.

Com as consultas em mãos, subimos o SCAC em três lojas do Comércio Eletrônico. Foram as lojas pertencentes aos nichos de Eletrodomésticos, Livraria e departamento. Para cada uma, avaliamos quais consultas foram mostradas e clicadas pelos usuários durante uma semana de teste. Os resultados foram compilados nas Figuras 4.2, 4.3 e 4.4.

Na Loja de Eletrodomésticos, 17 mil consultas foram mostradas para os usuários, enquanto que 11 mil delas foram clicadas. Das que foram mostradas, 94% vieram de consultas já realizadas anteriormente, e 40% foram extraídas do catálogo de produtos. Já para as consultas clicadas, 75% já tinham sido clicadas anteriormente e 59% vieram da base de produtos.

É possível notar que, para essa loja do segmento de eletrodomésticos, as consultas

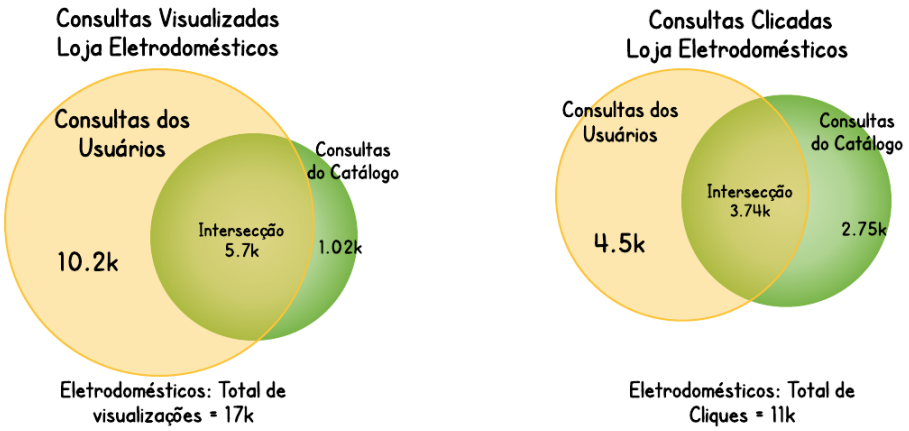


Figura 4.2. Distribuição das Consultas Mostradas e Clicadas para as duas fontes na Loja de Eletrodomésticos.

Tabela 4.6. Tabela com a quantidade de consultas mostradas e clicadas durante uma semana em 3 lojas de E-Commerce.

Loja	Consultas Mostradas	Consultas Clicadas
Eletrodomésticos	17 mil	11 mil
Livraria	160 mil	90 mil
Departamento	275 mil	118 mil

anteriores dos usuários são mais importantes que consultas extraídas do catálogo de produtos, tanto para mostrar aos usuários, quanto na participação em cliques: 41% dos cliques vieram de consultas que foram extraídas somente por essa fonte. Mesmo assim, não é bom deixar de extrair consultas do catálogo já que elas também tiveram cliques. Também é interessante notar que das 17 mil consultas mostradas, apenas 40% foram extraídas pelo catálogo de produtos e, dessas, apenas 6% foram extraídas somente por essa fonte, já que as outras também estavam nas consultas anteriores dos usuários. Quando olhamos os cliques, percebemos que a participação do catálogo de produtos é maior em relação às consultas mostradas: 25% das consultas clicadas vieram somente por essa fonte, o que indica uma boa importância, apesar de menor que o comportamento dos usuários.

Vimos na Seção 3.2 que a geração de consultas a partir do catálogo de produtos tem custo em função do tamanho do catálogo de produtos, enquanto que consultas anteriores dos usuários o custo é em função da quantidade de acessos. Para essa loja, vimos, pela Tabela 4.1, que tanto o catálogo de produtos quanto a quantidade de acessos representam valores baixos, ou seja, não há problemas de desempenho decorrentes do

uso das duas fontes simultaneamente, mas vale ressaltar que consultas anteriores dos usuários, especificamente para essa loja, têm mais utilidade para o sistema. Não é possível generalizar que estes resultados serão repetidos ou parecidos para outras lojas do mesmo segmento, dado que a análise foi feita considerando apenas uma loja.

A Loja do Segmento de Livraria teve 160 mil consultas mostradas e 90 mil dessas foram clicadas. A distribuição entre as fontes foi um pouco diferente da distribuição vista na loja de eletrodomésticos. A loja do segmento de livraria tem um catálogo muito grande, mas a maior participação tanto de consultas mostradas, quanto de clicadas foi da fonte das consultas anteriores dos usuários: 90% das consultas mostradas e 97% das clicadas vieram dessa fonte.

Como essa loja possui muitos produtos em seu catálogo, o processo de geração de consultas por essa fonte é bastante custoso em tempo de processamento. Sabendo-se que a maioria das consultas mostradas e clicadas pelos usuários são de origem de buscas anteriores dos usuários, retirar o catálogo como fonte geradora de consultas passa a ser uma boa opção, já que deixaria o sistema de geração bem mais rápido e sem perder tanto em qualidade.

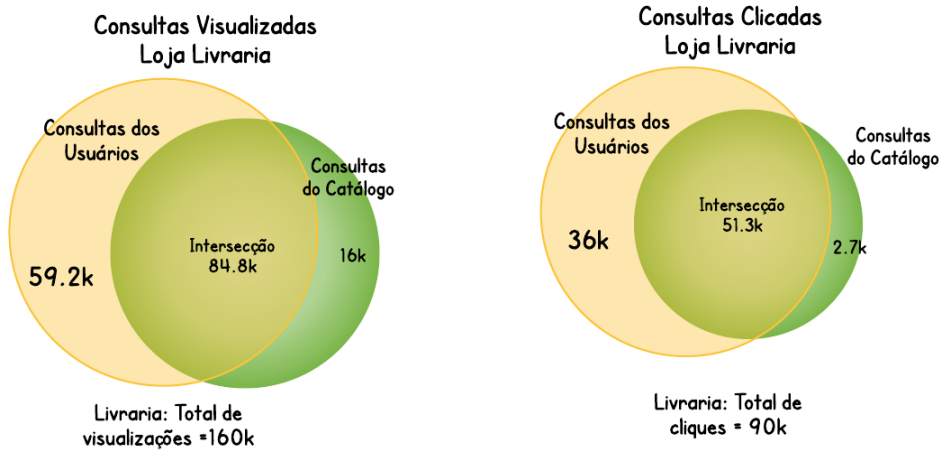


Figura 4.3. Distribuição das Consultas Mostradas e Clicadas para as duas fontes na Loja de Livraria.

Como pode ser visto na Tabela 4.6, a Loja do Segmento de Departamentos tem características peculiares. Essa loja vende produtos de diversas categorias diferentes e essa loja, em específico, atende muitos usuários. Nessa loja, em uma semana, o SCAC apresentou 275 mil consultas, enquanto que os usuários clicaram em 118 mil delas.

Pela Figura 4.4, vemos que 95% das consultas mostradas para os usuários pertencem à fonte de consultas já feitas anteriomente, enquanto que apenas 56% das consultas

mostradas advêm do catálogo de produtos. Das consultas clicadas, 97% eram consultas anteriores dos usuários e 75% são provenientes do catálogo, mas essa fonte contribuiu com apenas 3% dos cliques, já que os demais também foram gerados pela outra fonte. Assim, para essa loja, também pode-se retirar o catálogo de produtos como fonte de geração de consultas que não terá muitas perdas de qualidade, mas deixará o sistema de geração bem mais rápido por se tratar de uma loja que também tem muitos produtos em seu catálogo.

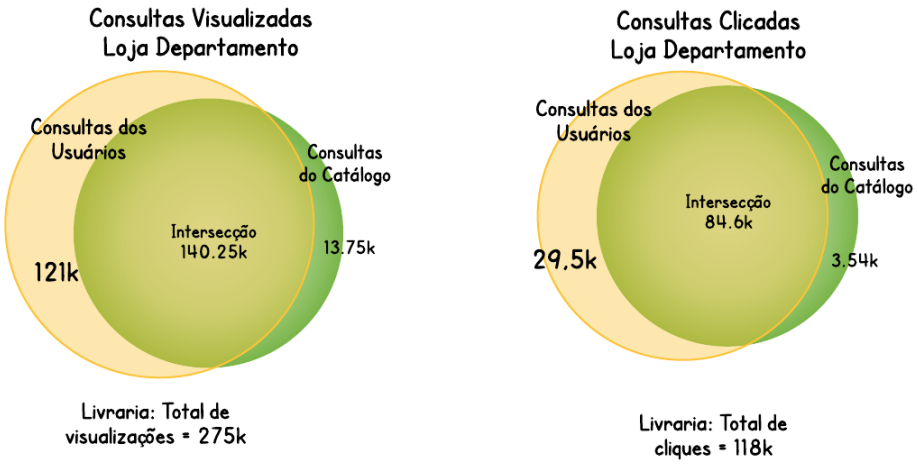


Figura 4.4. Distribuição das Consultas Mostradas e Clicadas para as duas fontes na Loja de Departamentos.

Com os resultados obtidos da marcação das consultas quanto à fonte geradora, percebemos que as consultas já realizadas anteriormente por usuários, para estas três lojas, são as melhores fontes de sugestões de complemento de consultas, ou seja, os usuários destas lojas acabam buscando por termos que já foram buscados por outros usuários anteriormente. O catálogo de produtos não se mostrou uma fonte de consultas promissora, apesar de ter sua importância, fazendo com que, dependendo da loja, seja mais prudente deixar de processá-lo e economizar em tempo de geração de consultas.

Na verdade, dependendo da configuração da loja, utilizar o catálogo de produtos pode ser crucial, pois contém informações que podem nunca terem sido buscadas, como nome de autor, livros e etc. Além disso, o comportamento de usuários pode não ser conhecido na maioria das lojas por ser uma fonte de difícil acesso para extração de informações.

Outra conclusão que podemos obter deste experimento é que, aparentemente, a forma como extraímos consultas do catálogo de produtos não é suficientemente boa e

devemos repensar como fazê-la. Outras estratégias de geração de consultas a partir desta fonte podem mudar os números obtidos neste experimento.

Em geral, se o lojista tem acesso às buscas anteriores de seus usuários, é extremamente importante utilizá-las como fonte de geração de consultas para SCAC. Se seu catálogo é grande, retirá-lo como fonte de geração de consultas é o melhor caminho a seguir para diminuir custos. Se o catálogo for pequeno, o melhor é utilizá-lo também na geração, já que essa fonte tem sua importância e não irá comprometer a loja em relação ao custo de processamento.

4.3.2 Explorando as fontes de Comportamento dos Usuários

Dentro dos sites de comércio eletrônico, os usuários interagem de várias formas com o sistema de busca. Usuários buscam produtos por algum termo, clicam em produtos de seu interesse e também os compram. Esta sequência de ações dos clientes nas lojas virtuais permite extrair diversas informações valiosas para serem usadas no SCAC. Geralmente, os trabalhos sobre SCAC na literatura possuem apenas as consultas feitas pelos usuários como base de teste. Nesta dissertação, temos acesso a todo esse processo de compra, permitindo o teste utilizando essas informações.

Esse rastreamento permite usarmos essas informações de diversas maneiras. Escolhemos duas dessas para explorar essa fonte. O primeiro estudo diz respeito a diferentes formas de atribuir importância para as consultas do SCAC quanto a origem da fonte e o segundo estudo trata sobre diferente variações no tempo para gerar a pontuação das consultas.

4.3.2.1 Origem da fonte de Comportamento dos Usuários

Com o rastreamento das interações dos usuários nos sites de comércio eletrônico, conseguimos gerar a sequencia de ações dos usuários. Com essas informações, conseguimos obter a quantidade de vezes que uma busca foi feita durante N dias (chamaremos de consultas em geral), a quantidade de vezes que uma busca levou o usuário a clicar em algum produto (chamaremos de consultas com cliques) e a quantidade de vezes que uma busca levou o usuário a comprar algum produto (chamaremos de consultas com compras. Essas são as três fontes de geração de consulta usando o comportamento dos usuários. Mostramos um exemplo das diferenças dessas quantidade na Tabela 3.9. O experimento, então, consiste em saber qual dessas fontes de Comportamento dos Usuários é a melhor para se usar para gerar consultas e pontuá-las no SCAC.

A estratégia de gerar pontuação utilizando buscas em geral é o que foi proposto por Bar-Yossef & Kraus [1], fazendo sempre a ressalva que o contexto do SCAC para E-

Commerce é diferente em relação à WEB. Utilizamos o buscas em geral como *baseline* neste trabalho. A estratégia de consultas com cliques também pode ser implementada no contexto da WEB, se tiver como extrair essa informação.

Utilizamos as lojas dos segmentos de Eletrodomésticos, Cosméticos e Moda para realizar esse experimento. Para cada loja, extraímos 90 dias de comportamento dos usuários para as três fontes de comportamento de usuários usadas e, assim, geramos as pontuações de cada consulta: consultas em geral, consultas com cliques e consultas com compras. Geramos a pontuação para cada consulta simplesmente contando a quantidade de ocorrências de cada uma em sua respectiva visão. Se uma consulta foi buscada três mil vezes, na visão de busca sua pontuação será de três mil. Se uma consulta levou um usuário a clicar em algum produto 25 vezes, a pontuação será essa e o mesmo ocorre para consultas que levaram o usuário a comprar. Para testar cada algoritmo, utilizamos as consultas feitas no dia seguinte aos 90 dias utilizados para a geração das pontuações. Por exemplo, se geramos as pontuações do dia 01 de Janeiro até 01 de Abril, as consultas para o teste foram tiradas do dia 02 de Abril. Isso deve ser feito para que não haja favorecimento ao algoritmo e para que seja respeitada a sequencia temporal dos fatos, o que sempre deve levado em conta quando tratamos de dados temporais.

Além disso, utilizamos a técnica de janela deslizante por 15 vezes. A janela deslizante faz com que o segundo teste, pelo exemplo acima, seja obtendo as consultas de 02 de Janeiro até 02 de Abril e o dia de teste seria o dia 03 de Abril. Isso é repetido 15 vezes, mudando as datas de início, fim e teste do experimento.

Para verificar qual estratégia é a melhor, utilizamos as métricas revocação a N (que chamaremos daqui em diante de R@N), com N valendo 3 ou 5, e o MRR. A R@3 verifica se a consulta que o usuário está buscando está entre as três primeiras alternativas de sugestões. Em todas essas métricas, quanto maior o valor obtido, melhor é o sistema. Também analisamos diferentes tamanhos do prefixo: 1, 3, 5 e 9 caracteres digitados. Isso significa que, quando o usuário digitou apenas uma letra, três, cinco ou nove, verificamos todas as três métricas relacionando a sugestão mostrada pelo sistema e a consulta que o usuário deseja para cada um desses tamanhos de prefixo. Se a consulta desejada tem tamanho menor que o tamanho do prefixo analisado, ela é ignorada. É o caso da busca por "cama"para o tamanho de prefixo 5 ou 9, por exemplo.

A Figura 4.5 apresenta os resultados obtidos para a loja do segmento de Eletrodomésticos e usando a métrica R@3. Na Figura 4.6 temos o resultado com R@5 e na Figura 4.7, mostramos o resultado com MRR. Podemos observar que em quase todos os resultados dessa loja, a geração e pontuação da consulta baseada em consultas com cliques foi a melhor em relação ao baseline (frequência) e em relação a consultas com

compras. Apenas para o tamanho do prefixo igual a um, no MRR, que consultas com compra se saiu melhor que a pontuação de consultas com clique, mesmo que por pouca diferença. A barra azul representa a fonte de consultas em geral, a barra vermelha representa consultas com cliques e a laranja, consultas com compra.

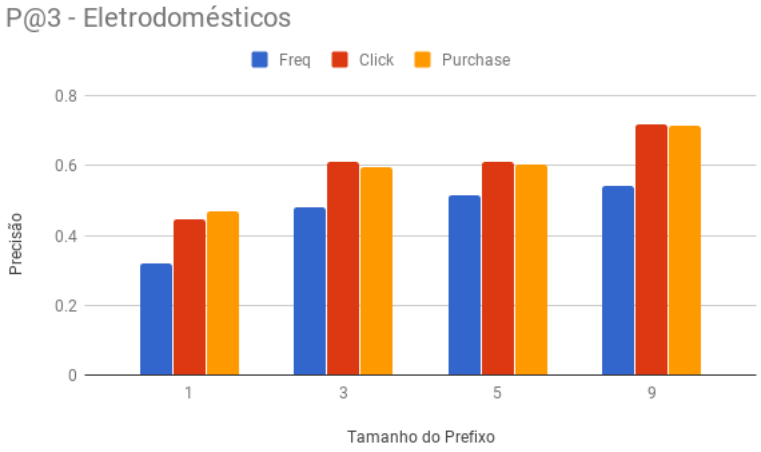


Figura 4.5. R@3 para a Loja do Segmento de Eletrodomésticos variando o tamanho do prefixo para as diferentes fontes de Comportamento do Usuário.

É interessante notarmos o quanto a fonte de consultas simples (baseline) tem desempenho bem pior do que as outras fontes nessa loja. Para a R@5, a fonte de consultas com cliques levou o sistema a atingir cerca de 67% de Revocação para o Tamanho de Prefixo igual a 3, um valor bastante alto para o SCAC.

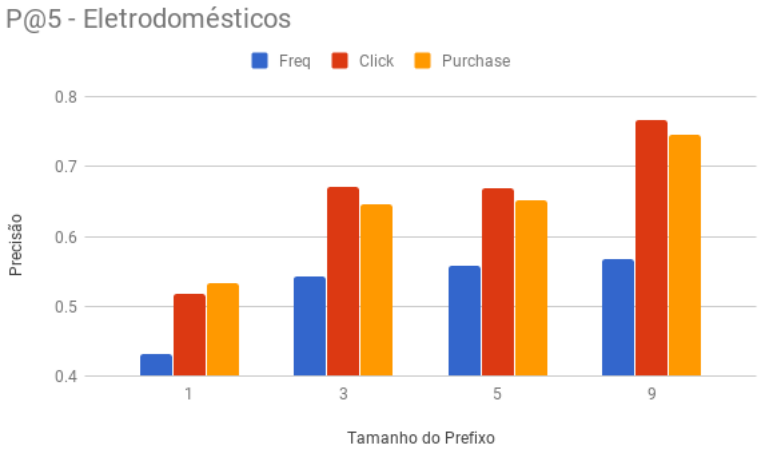


Figura 4.6. R@5 para a Loja do Segmento de Eletrodomésticos variando o tamanho do prefixo para as diferentes fontes de Comportamento do Usuário.

A Loja de Eletrodomésticos tem características únicas como catálogo pequeno e pouca diversidade nos produtos. Além disso, é a loja que teve menos usuários repetindo as mesmas consultas em uma sessão (Ver Tabela 4.5). Portanto, é uma das lojas mais prováveis que tenha um índice de desempenho de Revocação e MRR menor que as demais, onde mais usuários buscam as mesmas coisas.

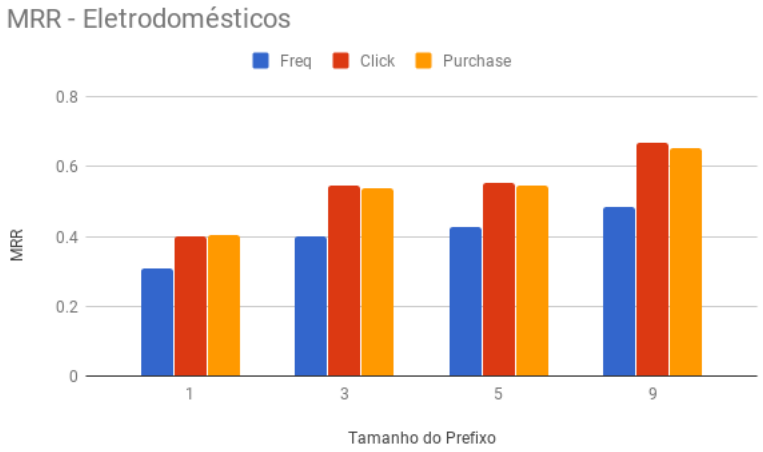


Figura 4.7. MRR para a Loja do Segmento de Eletrodomésticos variando o tamanho do prefixo para as diferentes fontes de Comportamento do Usuário.

A R@5 na loja de eletrodomésticos teve a pontuação de consultas com clique como melhor fonte de informação para prefixos com mais de três caracteres. Há um ganho de 23% em relação ao baseline e 4% em relação à informação de buscas que levaram a compra. Já quando olhamos nove caracteres, o aumento de revocação que consultas com cliques supera a buscas em geral é de 35% e consultas com compras supera buscas em geral em 31%.

O maior ganho nesta loja foi na métrica R@3 para prefixos de tamanho igual a um. Neste cenário, consultas com cliques supera o *baseline* em 39% enquanto que consultas com compras o superou em 45%.

Na média dos tamanhos de prefixo, a R@3, R@5 e MRR de consultas com cliques em relação ao baseline (apenas buscas) obtiveram 28%, 25% e 34% de ganho, respectivamente. Enquanto que consultas com compras, também em relação ao baseline, tiveram 28%, 23% e 33% de ganho.

Para o segmento de Cosméticos, os resultados obtidos estão nas Figuras 4.8, 4.9 e 4.10, referentes a R@3, R@5 e MRR, respectivamente. Percebemos que para o Tamanho do Prefixo igual a um, a fonte de consultas com compras teve um rendimento melhor que as demais, mas foi superada pela fonte de consultas com cliques para os

demais tamanhos de prefixos. Assim como na Loja de Eletrodomésticos, o clique foi o que mais sobressaiu positivamente nos resultados das três métricas.

Uma característica bastante peculiar da loja do segmento de cosméticos é que ela tem a menor média de termos e letras digitadas pelos usuários, fazendo com que as métricas de tamanho de prefixos 3, 5 e 9 tivessem um resultado bastante semelhantes entre si.

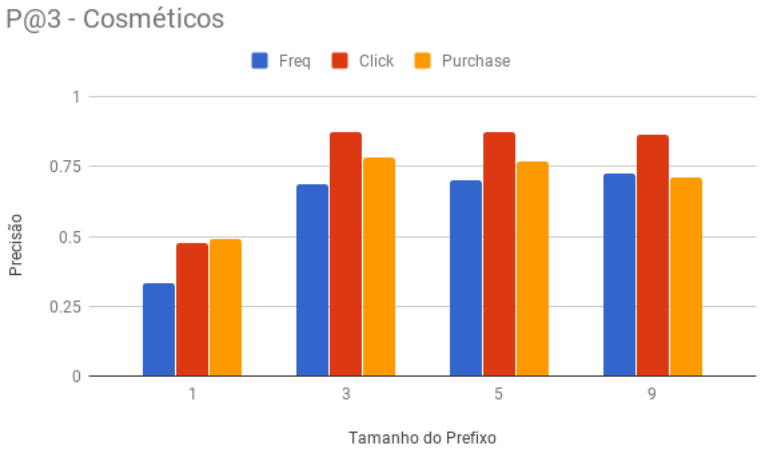


Figura 4.8. R@3 para a Loja do Segmento de Cosméticos variando o tamanho do prefixo para as diferentes fontes de Comportamento do Usuário.

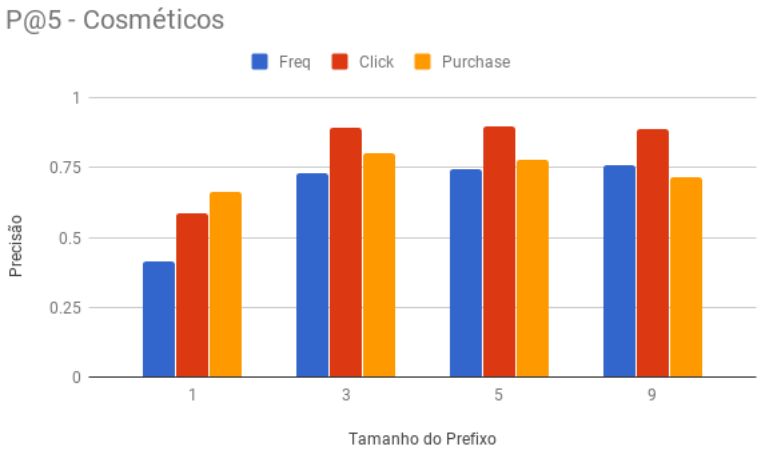


Figura 4.9. R@5 para a Loja do Segmento de Cosméticos variando o tamanho do prefixo para as diferentes fontes de Comportamento do Usuário.

Na loja do segmento de cosméticos, o ganho das consultas com cliques tem um peso mais evidente ainda que o obtido na loja de Eletrodomésticos. A R@5 para o

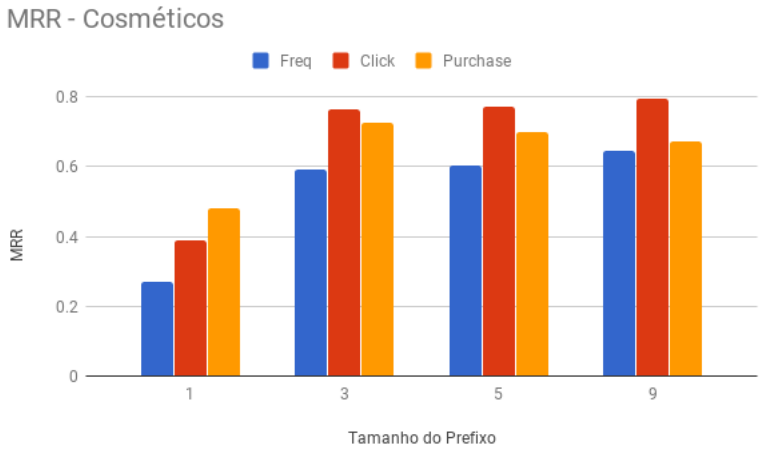


Figura 4.10. MRR para a Loja do Segmento de Cosméticos variando o tamanho do prefixo para as diferentes fontes de Comportamento do Usuário.

Tamanho de Prefixo igual a nove gerou um aumento de quase 17% em relação a usar somente consultas em geral, e 23% em relação a usar informação de consultas com compras. A partir do tamanho de prefixo igual a três, a R@5 chegou a atingir quase 90% para a estratégia de consultas com cliques, ou seja, a cada dez consultas no SCAC, a partir da terceira letra, em nove a sugestão que o usuário deseja estava exposta nas cinco primeiras sugestões. Neste cenário, buscas com cliques superou buscas em geral em 22%.

Na média dos tamanhos de prefixo, a R@3, R@5 e MRR de consultas com cliques em relação ao baseline (apenas buscas) para a loja do segmento de cosméticos obtiveram 26%, 23% e 29% de ganho, respectivamente. Enquanto que consultas com compras, também em relação ao baseline, tiveram 12%, 12% e 22% de ganho para esta loja.

A loja do segmento de Moda foi a única onde a frequência de consultas em geral como fonte de pontuação da consulta foi a melhor em comparação com consultas com cliques e consultas com compras. Os resultados dessa loja estão nas Figuras 4.11, 4.12 e 4.13 para R@3, R@5 e MRR, respectivamente.

Essa loja tem algumas características bem comuns em relação à loja de cosméticos quando falamos de tamanho do catálogo, quantidade de usuários e diferentes tipos de categorias de produtos vendidos. A diferença está principalmente no tamanho dos termos buscados pelos usuários. Enquanto que a loja de cosméticos tem a menor média deste tamanho, a loja de Moda possui uma das médias mais altas. Essa loja também possui a taxa de repetição de consulta mais alta de todas que observamos, chegando a 59% dos usuários buscarem mais de uma vez a mesma coisa.

Apesar de tudo isso, a Loja de Moda foi a que apresentou a menor taxa de R@3,

R@5 e MRR em comparação com as demais, chegando a apenas 0.39 de MRR para o Tamanho de prefixo igual a 9, considerado um valor bem baixo para esse tamanho, onde o usuário quase que digitou tudo o que queria.

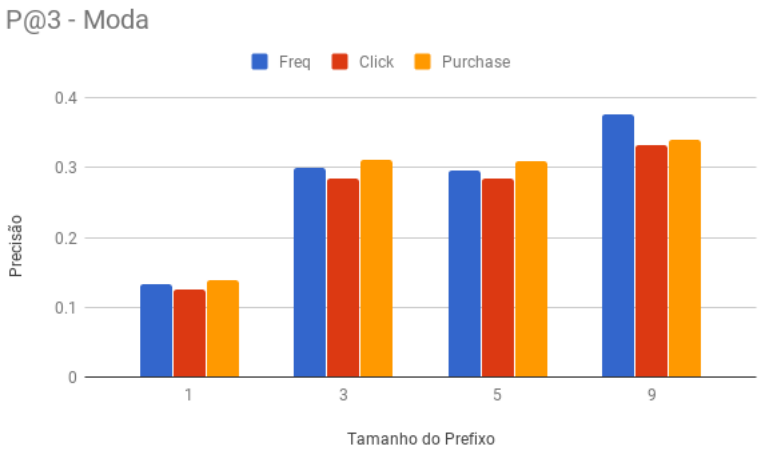


Figura 4.11. R@3 para a Loja do Segmento de Moda variando o tamanho do prefixo para as diferentes fontes de Comportamento do Usuário.

Figura 4.12. R@5 para a Loja do Segmento de Moda variando o tamanho do prefixo para as diferentes fontes de Comportamento do Usuário.

Na média dos tamanhos de prefixo, a R@3, R@5 e MRR de consultas com cliques em relação ao baseline (apenas buscas) para a loja do segmento de cosméticos obtiveram uma redução de -7%, -8% e -9% de perda, respectivamente. Enquanto que consultas com compras, também em relação ao baseline, tiveram -1%, -1% e -2% de perda para esta loja.

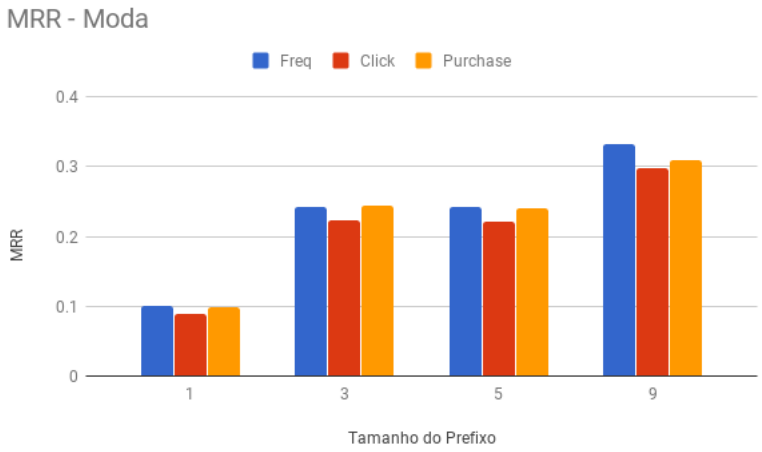


Figura 4.13. MRR para a Loja do Segmento de Moda variando o tamanho do prefixo para as diferentes fontes de Comportamento do Usuário.

Os experimentos indicam que não se pode apontar uma única fonte como sendo a melhor para os cenários estudados. Em dois segmentos, a pontuação gerada pelas consultas com cliques foi a que obteve melhor desempenho, enquanto que no outro segmento, a simples contagem de buscas em geral ganhou das demais.

Isso não quer dizer que as demais fontes não são importantes, pois realizamos os testes aplicando apenas uma fonte de cada vez e em apenas três lojas do e-commerce. É provável que, combinando as fontes, tenhamos resultados diferentes dos que obtivemos aqui. Deixamos esses outros experimentos e análises para os trabalhos futuros.

4.3.2.2 Variações do tempo para extrair informações da fonte de Comportamento dos Usuários

Outro fator que estudamos foi a variação do tempo usado para extrair informações da fonte de comportamento dos usuários e como ela impacta na performance do SCAC. Como vimos na Seção 4.3.2, a fonte de comportamento dos usuários que tem melhor desempenho na maioria dos segmentos testados foi consultas com cliques. Por isso, nesta Seção avaliamos apenas essa fonte geradora.

Utilizamos para o experimento as três lojas do experimento anterior, ou seja, os segmentos de Eletrodomésticos, Cosméticos e Moda. O experimento consiste em variar a quantidade de dias utilizados para gerar a pontuação de consultas com cliques. A escolha dos dias para a variação segue o que foi testado no trabalho de Whiting & Jose [12]. Nesse trabalho, o autor utilizou a janela de 2, 4, 7, 14 e 28. Mas também acrescentamos a quantidade de 90 dias de comportamento dos usuários. Chamamos

essas estratégias de *Most Popular Clicks Window N* ou mpc-w-N, onde N é a quantidade de dias utilizados. Utilizamos essa mesma regra para tanto para pontuar quanto para gerar as consultas, ou seja, para a janela de 2 dias, geramos apenas as consultas que levaram usuários a clicarem nesses dois dias, penalizando assim, consultas que não tiveram cliques.

Para a loja do segmento de eletrodomésticos, mostramos a R@3, R@5 e MRR nas figuras 4.14, 4.15 e 4.16, respectivamente.

Percebemos, nos resultados, que para essa loja, uma quantidade menor de dias de comportamento dos usuários já é suficente para atingir bons resultados de Revocação e MRR. No MRR para prefixos de tamanho 5 ou mais, é possível atingir mais de 0,75 no SCAC para essa loja, um valor bem significativo.

Também é possível notar que ter uma pontuação para muito tempo (90 dias, por exemplo) faz o desempenho do sistema piorar, levando a pensarmos que é uma loja onde os conceitos de novidade e tendência são bastante presentes, apesar de ser uma loja com pouca variedade de categorias e produtos. Como já dissemos anteriormente, esses resultados valem apenas para a loja em questão e não é possível generalizá-los nem para o segmento específico e muito menos para qualquer loja do E-Commerce.

Esta mudança no comportamento dos usuários ao longo do tempo nos leva a pensar que os usuários dessa loja buscam por coisas diferentes ao longo do tempo. Um caso bastante conhecido pelos lojistas é a época de verão, quando usuários buscam muito por aparelhos de ar condicionado e ventiladores, diferentemente da época do inverno, por exemplo.

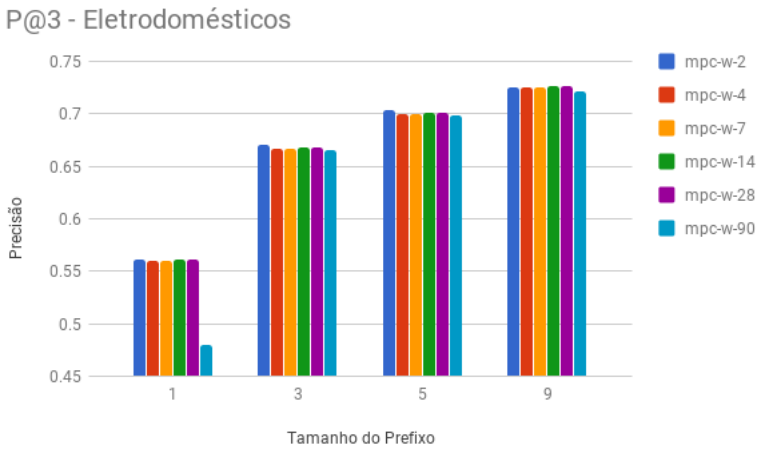


Figura 4.14. Resultado da R@3 para a Loja do Segmento de Eletrodoméstico variando o tamanho do prefixo para as diferentes quantidade de dias de Comportamento do Usuário.

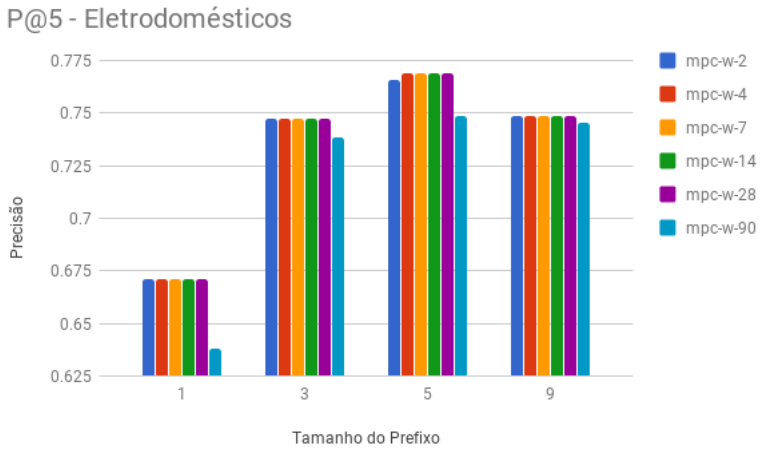


Figura 4.15. Resultado da R@5 para a Loja do Segmento de Eletrodomésticos variando o tamanho do prefixo para as diferentes quantidade de dias de Comportamento do Usuário.

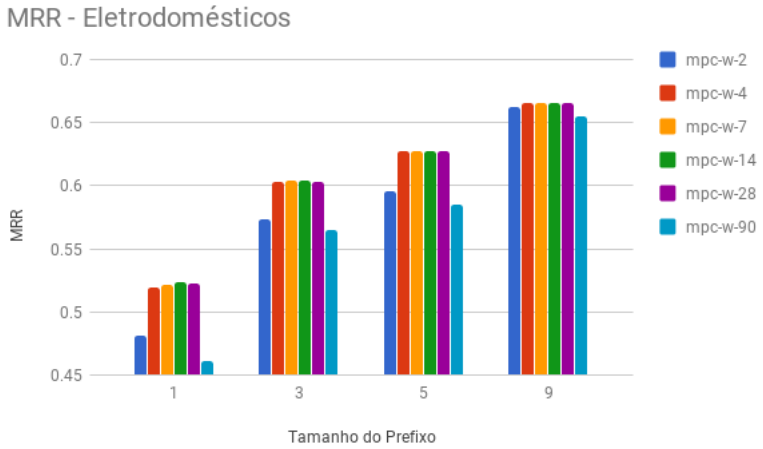


Figura 4.16. Resultado do MRR para a Loja do Segmento de Eletrodomésticos variando o tamanho do prefixo para as diferentes quantidade de dias de Comportamento do Usuário.

As Figuras 4.17, 4.18 e 4.19 mostram os resultados para a loja do segmento de cosméticos com R@3, R@5 e MRR mostrados, respectivamente. Nessa loja, não conseguimos gerar relevância estatística para concluirmos se uma quantidade de dias é melhor que outras.

Em valores absolutos, essa loja atingiu pontuações de revocação e MRR elevadíssimos e podemos concluir que não há muita variação no comportamento dos usuários ao longo do tempo. Muito provavelmente porque as marcas vendidas já são consolidadas no mercado e não deve haver muita variação na concorrência.

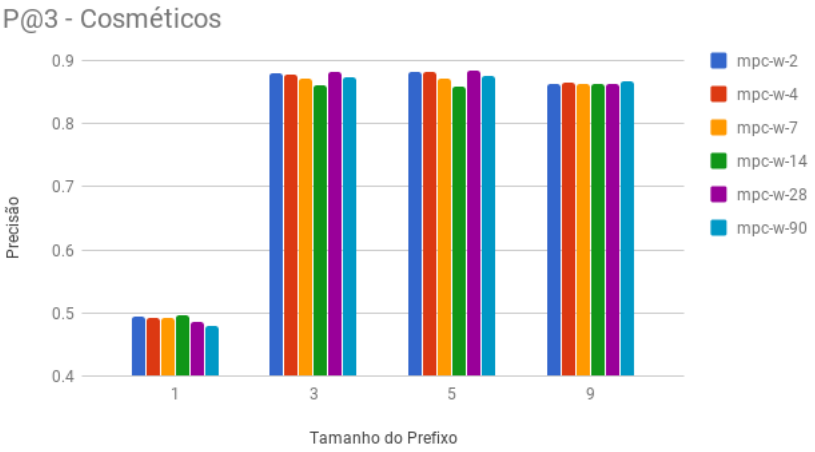


Figura 4.17. Resultado da R@3 para a Loja do Segmento de Cosméticos variando o tamanho do prefixo para as diferentes quantidade de dias de Comportamento do Usuário.

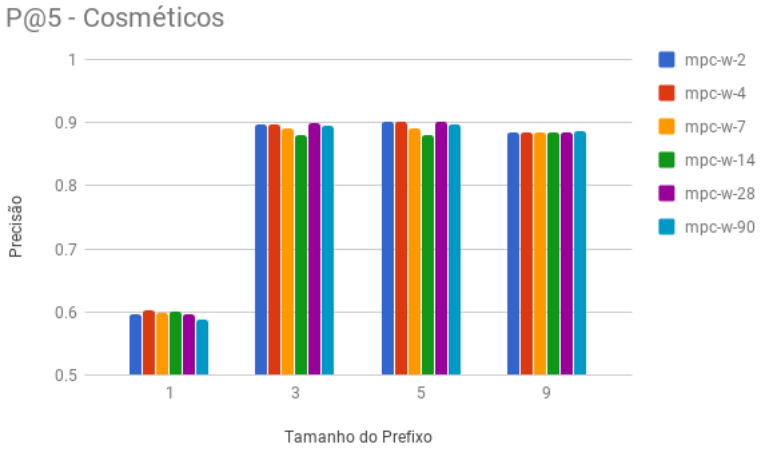


Figura 4.18. Resultado da R@5 para a Loja do Segmento de Cosméticos variando o tamanho do prefixo para as diferentes quantidade de dias de Comportamento do Usuário.

As Figuras 4.20, 4.21 e 4.22, apresentam os resultados de p@3, p@5 e MRR, respectivamente, para a loja do segmento de moda. Essa loja teve resultados bem diferentes dos demais segmentos analisados anteriormente. Os resultados revelam que, para essa loja, quanto maior a quantidade de dias coletados do comportamentos dos usuários para pontuar a consulta, melhor é o desempenho do sistema. Em valores absolutos, esse segmento não atinge bons desempenho, deixando uma lacuna grande para melhorar as sugestões de consultas.

Os resultados gerados para as três lojas de segmentos diferentes não foram se-

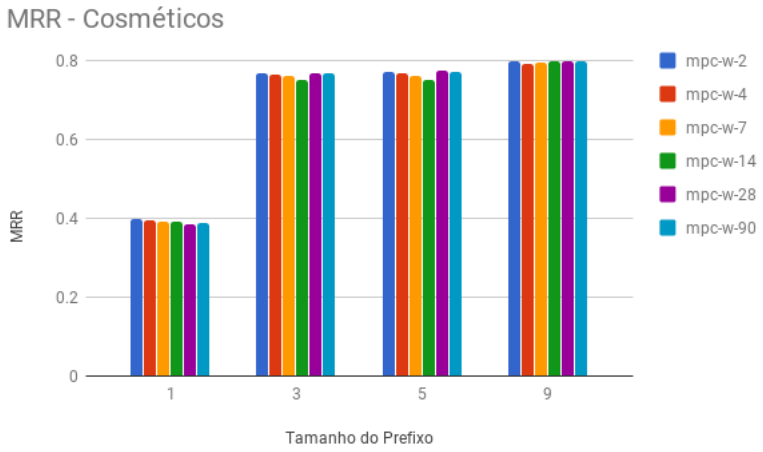


Figura 4.19. Resultado do MRR para a Loja do Segmento de Cosméticos variando o tamanho do prefixo para as diferentes quantidade de dias de Comportamento do Usuário.

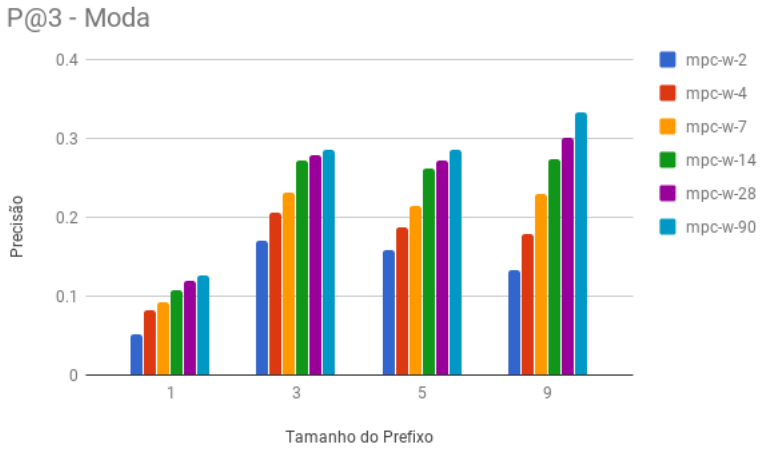


Figura 4.20. Resultado da R@3 para a Loja do Segmento de Moda variando o tamanho do prefixo para as diferentes quantidade de dias de Comportamento do Usuário.

melhantes entre si. Na loja de Eletrodomésticos, poucos dias de comportamento dos usuários já foram suficientes para ter um bom desempenho do SCAC. Na loja de Cosméticos, a quantidade de dias não fez diferença significante no resultado e na loja de Moda, mais tempo de extração do comportamento leva a melhor desempenho.

Com esses resultados obtidos, é provável que a quantidade de dias para extrair informações dos usuários esteja ligada a uma função que relaciona a quantidade de acessos que uma loja possui com o tamanho do catálogo e a diversidade dos produtos. Pela Tabela 4.1, a loja do segmento de eletrodomésticos possuem poucos acessos, mas

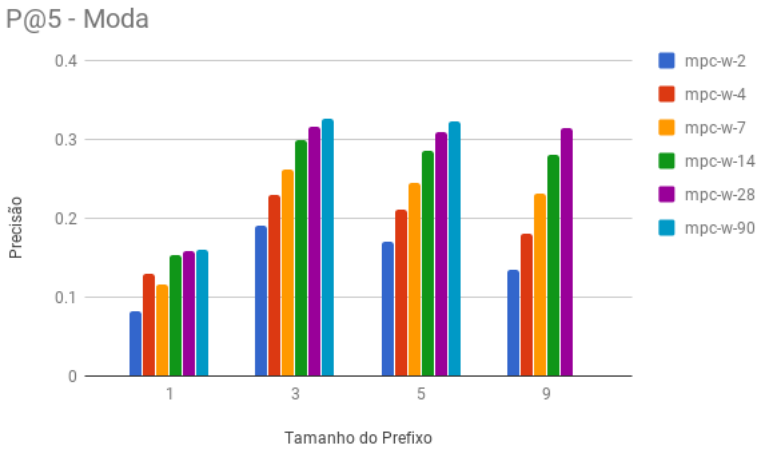


Figura 4.21. Resultado da R@5 para a Loja do Segmento de Moda variando o tamanho do prefixo para as diferentes quantidade de dias de Comportamento do Usuário.

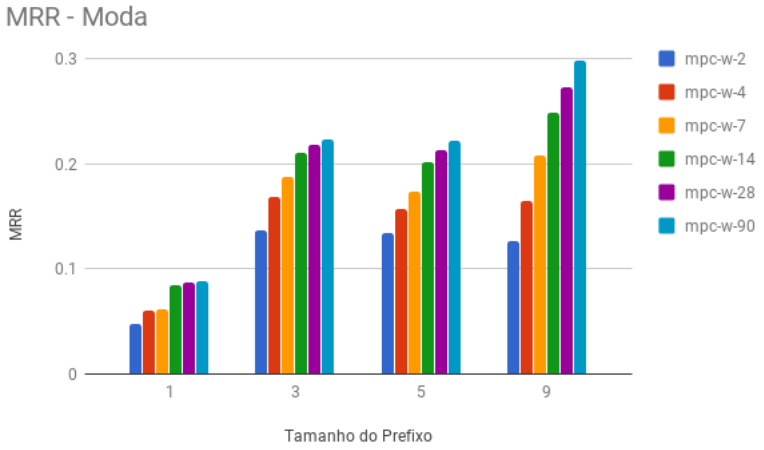


Figura 4.22. Resultado do MRR para a Loja do Segmento de Moda variando o tamanho do prefixo para as diferentes quantidade de dias de Comportamento do Usuário.

também possui o catálogo pequeno e pouca diversidade de produtos. Isso pode indicar que poucos dias de informações dos usuários já são suficientes para o SCAC aprender a pontuar e ordenas as consultas dessa loja. O mesmo pode ser observado para a loja de Cosméticos. Ela tem bastante acesso e o tamanho do catálogo é pequeno assim como a diversidade de produtos, ou seja, poucos dias de carga de informação são suficientes para o SCAC. Na loja de Moda, em comparação com a loja do segmento de Cosméticos, a quantidade de acessos é menor e o tamanho do catálogo, maior. Nesse segmento, precisou-se de mais dias para o SCAC aprender a ordenar melhor as

consultas, também podendo indicar a relação citada.

$$Perf_{SCAC} = \frac{Quantidade\_Acessos}{Tamanho\_Catalogo * Diversidade\_Produtos}$$

A fórmula acima tenta traduzir o que dissemos no parágrafo anterior. A *Performance* do SCAC é proporcional a quantidade de acessos, mas inversamente proporcional ao tamanho do catálogo e diversidade de produtos. Se tivermos a tabela 4.1 como referência, conseguimos os seguintes números para as três lojas testadas:

$$Perf_{Eletrodomstico} = 5$$

$$Perf_{Cosmticos} = 2.23$$

$$Perf_{Moda} = 0.25$$

Esses valores condiz com o resultado obtido neste experimento, ou seja, se a loja possui muitos produtos e uma boa diversidade, será preciso muitos acessos para que o SCAC atinja uma boa *performance.*

Concluímos, portanto, que, para as lojas experimentadas, a quantidade acessos, tamanho do catálogo e diversidade de produtos se relacionam para mostrar a quantidade de dias que deve ser usada para coletar informações dos usuários em lojas em Comércio Eletrônico. Quanto maior o tamanho do catálogo e diversidade de produtos, mais informações são necessárias coletar para o SCAC aprender a pontuar as sugestões de consulta. Se a loja já possui grandes quantidades de acessos diários, não são necessários extrair muitos dias para obter informações de entrada para o SCAC.

Outra ponto importante que percebemos nesse experimento é que o desempenho do SCAC já não melhora muito após a terceira letra digitada pelo usuário. Isso provavelmente se deve ao fato de que, após digitar três letras, a variedade de palavras no universo dessas lojas se reduz bastante. O prefixo "gel"em na loja de eletrodoméstico, por exemplo, provavelmente indica que o usuário quer buscar algo relacionado com "geladeira". Esse comportamento deve ser mais evidente quanto menor é a variedade de produtos que a loja vende.

Capítulo 5

Conclusão

O Sistema de Complemento Automático de Consulta (SCAC) é um dos primeiros sistemas com o qual um usuário interage em uma loja de comércio eletrônico. Saber quais consultas sugerir para cada usuário é uma tarefa desafiadora. Neste trabalho, mostramos algumas formas de extrair as consultas que devem ser apresentadas para esses usuários e estudamos qual fonte de consultas traz benefícios melhor para a loja. Além disso, exploramos o comportamento dos usuário para identificar qual é a melhor estratégia de ordenação para sugestões de busca.

Neste trabalho, propomos estudar quais das fontes geradoras de consultas para SCAC em Comércio Eletrônico é a mais importante para os usuários. Também estudamos a fundo a fonte do comportamento dos usuários e fizemos experimentos com diferentes formas de extrair informações do comportamento e diferentes conjuntos temporais.

Sobre as fontes geradoras de consultas, estudamos formas de gerar consultas pelo catálogo de produtos e pelo comportamento dos usuários. No catálogo de produtos, extraímos n-gramas a partir de campos estrurados como nome do produto, categoria, descrição e atributos. Após esse processo de extração, passamos as consultas por um filtro para não deixar "consultas ruins"irem para o SCAC. Por fim, validamos se essas consultas trariam resultados no Sistema de Busca antes de liberá-las para serem usadas pelo SCAC. Para o Comportamento dos Usuários, pegamos as consultas que foram realizadas pelos usuários nos últimos N dias.

No estudo mais profundo sobre o Comportamento dos Usuários, experimentamos a diferença de usar as consultas feitas pelos usuários, usar consultas que leveram usuários a clicarem em algum produto e usar consultas que leveram usuários a comprar algum produto. Além desse experimento, também verificamos se há diferença entre variar quantidade de dias usados para obter essas informações em 2, 4, 7, 14, 28 e 90

dias.

Como resultado do primeiro experimento, vimos que as consultas geradas a partir do Catálogo de Produtos não tiveram tantas visualizações e cliques quanto as consultas geradas pelo Comportamento dos Usuários. Vimos também que o tempo de processamento para a geração de consultas pelo catálogo cresce de acordo com o tamanho deste. Também percebemos que o gerador que fizemos não é capaz de gerar boas consultas para o SCAC. Por outro lado, a informação sobre as consultas que os usuários já fizeram em dias anteriores conseguiu atingir cerca de 95% do total de visualizações, e 97% do total de cliques em uma loja real do segmento de Departamentos. Nas outras duas lojas do segmento de Livraria e Eletrodomésticos, as porcentagens de consultas visualizadas a partir do Comportamento dos Usuários foram de 90% e 94%, respectivamente. Para consultas clicadas, as taxas foram de 97% e 75%.

O experimento sobre o Comportamento dos Usuários foi feito em três lojas de segmentos diferentes: Eletrodomésticos, Cosméticos e Moda. Nos segmentos de Eletrodomésticos e Cosméticos, usar consultas que levaram os usuários a clicar em algum produto foi a melhor estratégia em relação ao baseline (frequência simples) e em relação às consultas com vendas que foi a segunda estratégia adotada. Para a loja do segmento de Moda, a frequência pura das consultas submetidas ao sistema de busca foi a que deu maior ganho de performance. Na loja de Cosméticos, por exemplo, utilizar consultas com clique gerou um ganho de 23% em relação ao baseline e 17% em relação a buscas com compras para a métrica R@5 com tamanho do prefixo igual a nove. Esse é um valor bem expressivo de ganho para essa métrica em SCAC para E-commerce.

Para o estudo de variações temporais na extração de informações do Comportamento dos Usuários, também utilizamos as três lojas dos segmentos de Eletrodomésticos, Cosméticos e Moda. Com os resultados obtidos, percebemos que, provavelmente, a melhor quantidade de dias para extrair informações é uma função que envolve a quantidade de acessos da loja, o tamanho do catálogo e a diversidade de produtos. A loja do segmento de Cosméticos que tem bastante acesso de usuários, tamanho do catálogo pequeno e pouca diversidade de produtos fez com que os resultados para os diferentes dias não se alterasse para as diferentes métricas que submetemos: Revocação a 3, Revocação a 5 e MRR. Já a loja de moda que tem o tamanho do catálogo maior, as três métricas mostraram que usar mais dias fazem o sistema melhorar.

Com os dados obtidos, percebemos que o Comportamento dos Usuários é bastante útil como fonte geradora de consultas para SCAC em Comércio Eletrônico. Essa fonte é mais importante, em muitos casos, do que o catálogo de produtos oferecido pelos lojistas, apesar desse último também ter sua importância nos cenários citados.

Dentro do Comportamento dos Usuários, conseguimos diferenciar as consultas

buscadas, daquelas que levam usuários a clicarem em algum produto e daquelas que levam usuários a comprarem produtos clicados. Com essa diferenciação percebemos que utilizar as consultas que levam usuários a clicarem na geração e pontuação das consultas foi uma estratégia que fez o desempenho de duas das três lojas testadas aumentar. O MRR da loja de Eletrodomésticos utilizando esse cenário, para o Tamanho de Prefixo igual a três, foi 2% maior que buscas que levaram usuários a comprar algum produto e 36% maior que apenas as buscas realizadas.

Quanto à variação de quantidade de dias utilizados na coleta das informações de buscas dos usuários, não há um ganho significativo para todos os segmentos. Nesse caso, o lojista precisa entender o funcionamento da sua loja para verificar qual é a melhor quantidade de dias na extração: é possível dar uma visão mais sobre tendências ou popularidade fixa no SCAC.

Além das estratégias de ordenação estudadas aqui, muitas outras podem ser realizados para que o SCAC melhore no cenário de E-Commerce. Podemos, por exemplo, combinar as três características que estudamos do comportamento dos usuários: buscas normais, buscas com clique e buscas com compras, dando prioridades diferentes para cada cenário e experimentando em diversas lojas. Na mesma linha, podemos combinar as variações dos dias usados para extrair informação e até utilizar previsão temporal para saber qual a melhor combinação de dias a ser usado.

Além de utilizar consultas baseado no comportamento dos usuários no site, também podemos utilizar dados individuais deles para melhorar o desempenho do sistema. Informações como idade, gênero, cor favorita, tamanho de roupa e calçados preferidos, poder de compra, dentre muitas outras podem ser usadas para que o SCAC ofereça sempre as melhores sugestões. Também podemos utilizar a página atual do usuário para diferenciar sugestões de consulta: se um usuário está na página de categoria de jogos de playstation, sugestões como "call of duty"podem ter maior relevância do que "câmera digital"ou "cafeteira"para o prefixo "ca"digitado por esse usuário.

Outras métricas das que usamos neste trabalho também podem ser utilizadas para saber qual estratégia de SCAC é a melhor. Algumas métricas são do mundo do comércio eletrônico como o Faturamento e a Conversão. O faturamento é medido pela soma dos valores dos produtos comprados, enquanto a conversão é uma taxa de compras por visita. Talvez métricas assim sejam mais interessantes de medir do que a precisão e MRR.

Por último, também pode-se explorar diversos algoritmos de combinação de evidências para gerar a ordenação final das consultas no SCAC. Neste trabalho utilizamos a combinação linear de evidências e árvore de decisão simples para obter a ordenação final. Algoritmos como Deep Learning, Redes Neurais, SVM, dentre outros podem

também ser explorados para ordenar consultas.

Por fim, os experimentos realizados neste trabalho são os pioneiros na questão do estudo do mundo do SCAC para E-Commerce. As análises e estudos realizados e mostrados são apenas um primeiro passo para a evolução dos Sistemas de Complemento Automático de Consultas e Lojas de Comércio Eletrônico.

Referências Bibliográficas

[1] Bar-Yossef, Z. & Kraus, N. (2011). Context-sensitive query auto-completion. Em *Proceedings of the 20th international conference on World wide web*, pp. 107--116. ACM.

[2] Burges, C.; Svore, K.; Bennett, P.; Pastusiak, A. & Wu, Q. (2011). Learning to rank using an ensemble of lambda-gradient models. Em *Proceedings of the Learning to Rank Challenge*, pp. 25--35.

[3] Cai, F. & de Rijke, M. (2016). Learning from homologous queries and semantically related terms for query auto completion. *Information Processing & Management*, 52(4):628--643.

[4] Cai, F.; De Rijke, M. et al. (2016). A survey of query auto completion in information retrieval. *Foundations and Trends® in Information Retrieval*, 10(4):273--363.

[5] Chaudhuri, S. & Kaushik, R. (2009). Extending autocompletion to tolerate errors. Em *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pp. 707--718. ACM.

[6] John, K. (2016). 7 search autocomplete best practices for ecommerce sites.

[7] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Em *Soviet physics doklady*, volume 10, pp. 707--710.

[8] Li, L.; Deng, H.; Dong, A.; Chang, Y.; Zha, H. & Baeza-Yates, R. (2015). Analyzing user's sequential behavior in query auto-completion via markov processes. Em *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 123--132. ACM.

[9] Mitra, B. & Craswell, N. (2015). Query auto-completion for rare prefixes. Em *Proceedings of the 24th ACM international on conference on information and knowledge management*, pp. 1755--1758. ACM.

[10] Rangel, J. (1999). Loja real x loja virtual.

[11] Shokouhi, M. (2013). Learning to personalize query auto-completion. Em *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 103--112. ACM.

[12] Whiting, S. & Jose, J. M. (2014). Recent and robust query auto-completion. Em *Proceedings of the 23rd international conference on World wide web*, pp. 971--982. ACM.