http://ieeexplore.ieee.org/document/7524413/?part=1

Abstract:

FOCUS, as the first step to automatically debug high search response time in search logs. 为运营商的进一步检查提供了一个很好的起点。

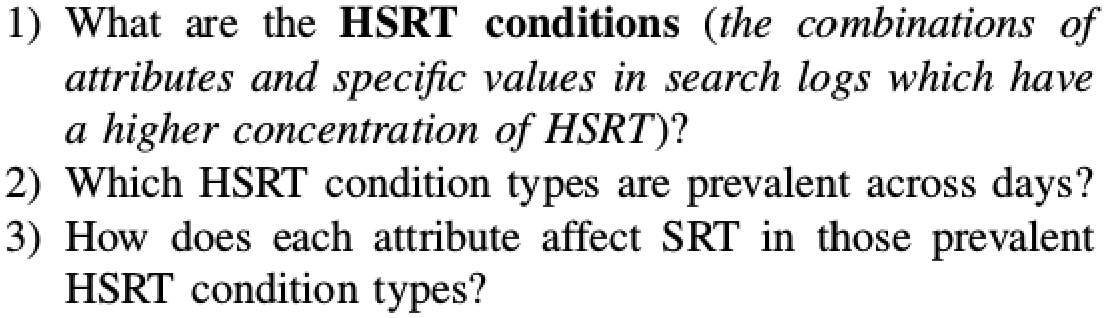
FOCUS 在百度上部署了 2.5 个月，并分析了 10 亿个搜索日志。与之前的方法（？？ critical clustering）相比，FOCUS 可以减少 90%的检查条目，同时具有更高的准确率和召回率。

图像密集搜索，及其分析优化。

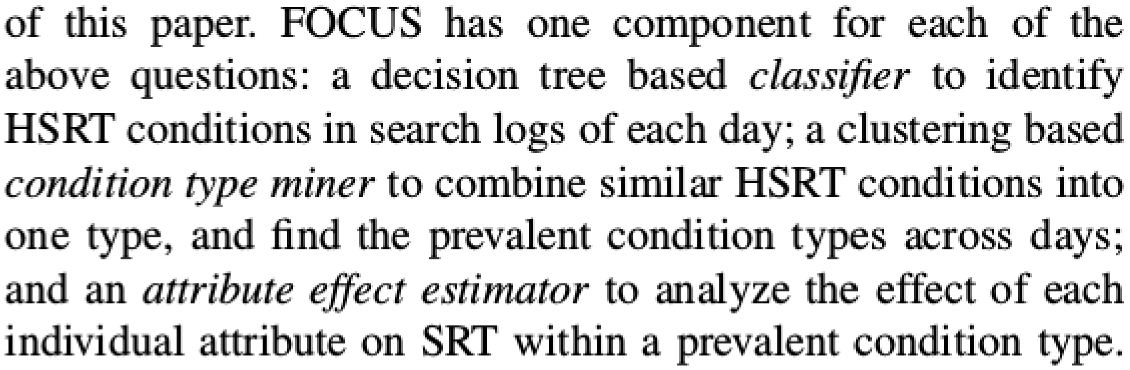
Introduction:

通常，搜索提供商将 javaScript 代理安装到搜索页面中，并在客户端测量用户感知的 SRT 和 SRT 组件。附加的*可观察的*和*可测量的*如果考虑到基于运算符的潜在影响 SRT，诸如嵌入图像的数量和页面是否包含 ADS 的属性也会被 JavaScript 代理记录，这些是基于运营团队的"领域知识"。

如摘要中说实现 FOCUS 自动化功能，那么应该能回答如下的问题：



1) What are the **HSRT conditions** (*the combinations of attributes and specific values in search logs which have a higher concentration of HSRT*)?
2) Which HSRT condition types are prevalent across days?
3) How does each attribute affect SRT in those prevalent HSRT condition types?



of this paper. FOCUS has one component for each of the above questions: a decision tree based *classifier* to identify HSRT conditions in search logs of each day; a clustering based *condition type miner* to combine similar HSRT conditions into one type, and find the prevalent condition types across days; and an *attribute effect estimator* to analyze the effect of each individual attribute on SRT within a prevalent condition type.

condition type miner？？

多维分类，即对 HSRT 的分类，目的分析出不同“异常”具体哪些属性的变化起到主导影响。

critical clustering？？ 决策树的分级机制 inflate 充气，膨胀

优化图像传输时延，现有方法： ### embedding base64-encoded images in HTML [10] 。×××A one-month real-world deployment (from day 44 to day 74) shows that the 80 th percentile of SRT has been reduced by 253ms, and the HSRT fraction has been reduced by one third.

Preliminaries:

搜集的属性被认为：potentially impact SRT based on the operators’ domain knowledge.

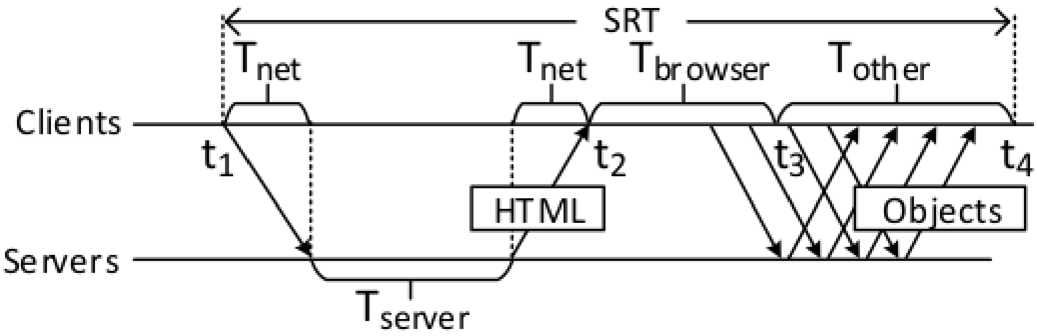


Fig. 1. Simplified timeline of a search.

意义：t1 浏览器提交查询；t2 浏览器接收并下载完 HTML；t3 浏览器完成 HTML 的解析；t4 页面渲染完成并最终呈现。

记录： Tserver 服务器 HTML 文件响应时间；
Tnet = t2 – t1 – Tserver 网络传输时间；
Tbrowser = t3 – t2 浏览器 HTML 解析时间；
Tother = t4 – t3 剩余处理时间。

识别 HRT 条件的挑战：

组合影响与依存关系；需要避免输出重叠的条件？？

Design

condition like {#images > 20, ads = yes}.

分类是识别空间中(非重叠)的判定边界的任务，并告诉哪个区域具有很大一部分的 HSRT。被识别为 HSRT 的那些区域可以用作 HRT 条件。

使用决策树作为分类器，

不使用关键聚类，避免识别条件重叠。J. Jiang, V. Sekar et al., "Shedding light on the structure of internet video quality problems in the wild", *CoNEXT*, 2013.

FOCUS 概况

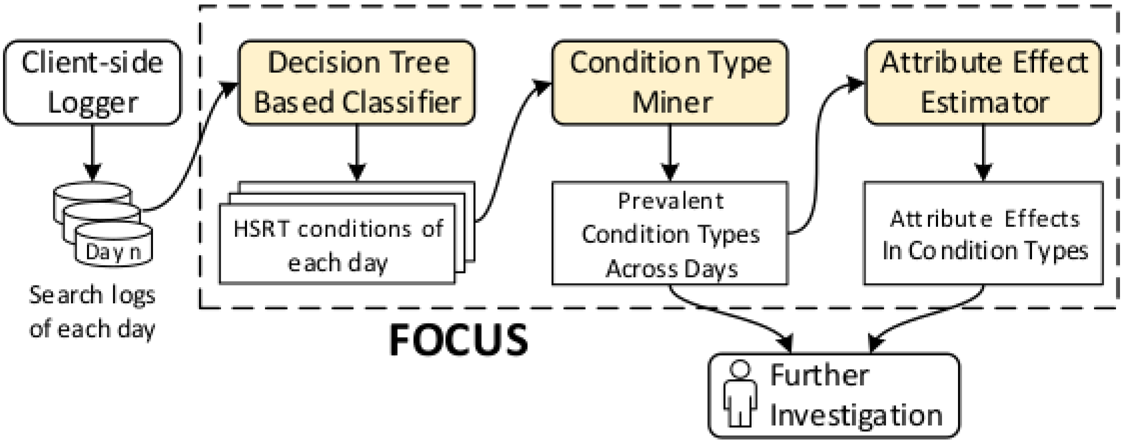

Client-side Logger
Day n
Search logs of each day
Decision Tree Based Classifier
HSRT conditions of each day
Condition Type Miner
Prevalent Condition Types Across Days
Attribute Effect Estimator
Attribute Effects In Condition Types
FOCUS
Further Investigation

Fig. 6. Overview of FOCUS.

第一

首先使用决策树分类器识别出搜索日志中的 HRT 条件

第二

使用 condition type miner 挖掘不同天搜索日志的类似的条件，从而归纳出常见的条件类型

第三

使用属性效应估计器分析常见条件类型中的属性如何影响 SRT 和 SRT 组件。这些流行的条件类型及其对 SRT 的属性效应，以焦点的形式输出，为操作者提供了进一步的调查的有价值的起点。

Decision Tree based Classifier

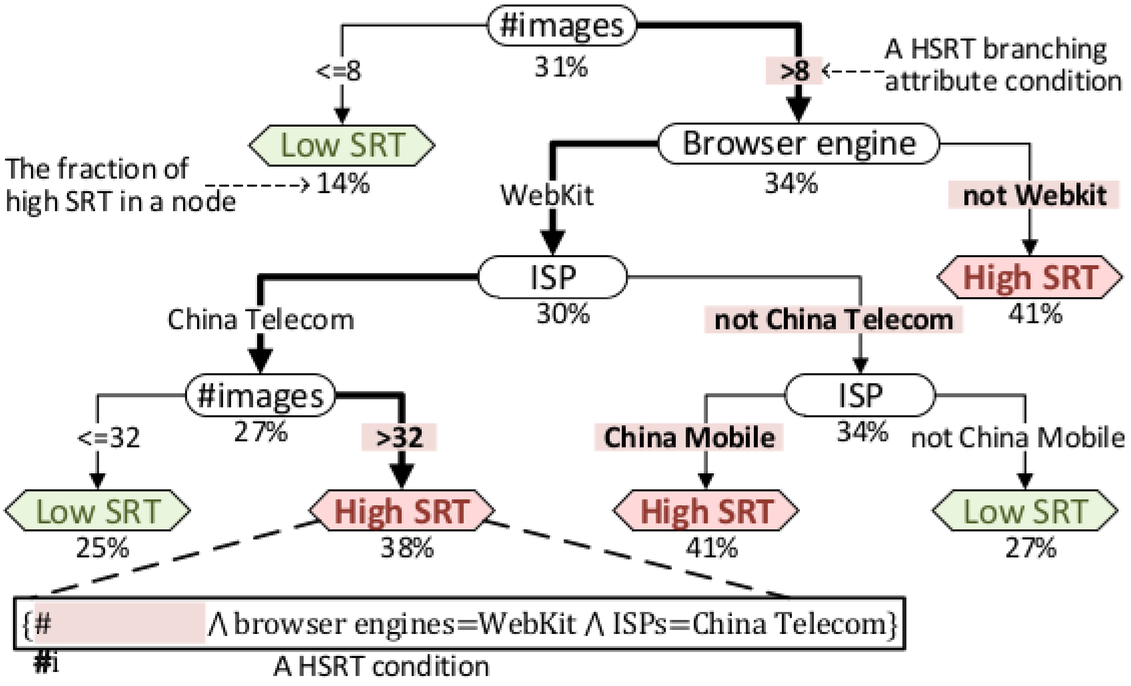


Fig. 7. Decision tree built from our one-day data. The thick path identifies a HSRT condition. HSRT branching attribute conditions are in bold and shaded.

Each path from the root node to a HSRT leaf node represents a HSRT condition, which is described by a logical AND (^) combination of the attribute conditions along the path.

***(??)

如何构建决策树以及我们在其中的特殊设计选择，以便我们可以使用决策树来识别我们想要的 HRT 条件：

（1）到（5）

Conditional Type Miner

具体地，一组 HRT 条件属于相同 **HSRT 条件类型**当且仅当它们具有

(i)属性的相同组合时，
(ii)每个类别属性的相同值，和
(iii)*相似的*每个数字属性的间隔。

对于第三个条件的量化度量：

To measure the similarity between clusters A and B, we use sim(A, B) = min{Jacc(a, b) : a 2 A, b 2 B}, where a and b are the attribute intervals in A and B, respectively.

Attribute Effect Estimator

Since the number of search logs of each day is huge, both C and C i 0 can contain enough data to show reliable statistics. As a result, we believe that the historical data based comparison can provide a reasonable estimate of the attribute effects, similar to the spirit of control experiments.

RESULTS

A. Evaluation

B. Condition Types

在本文的其他部分，我们关注流行的条件类型（在一个月内出现超过 5 天）。

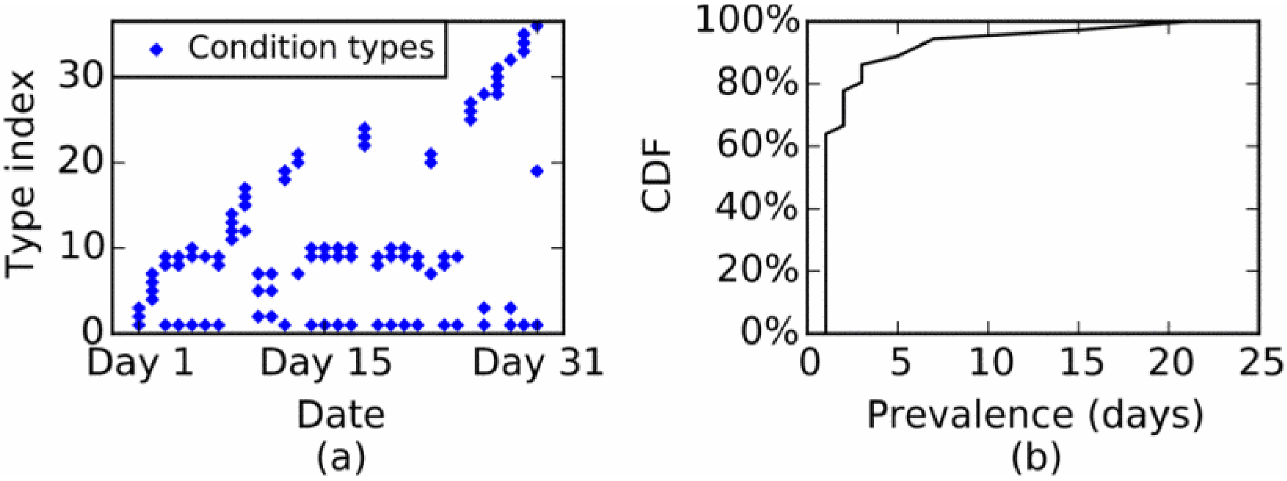
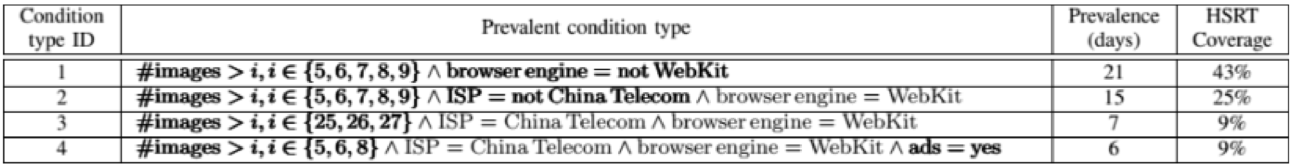
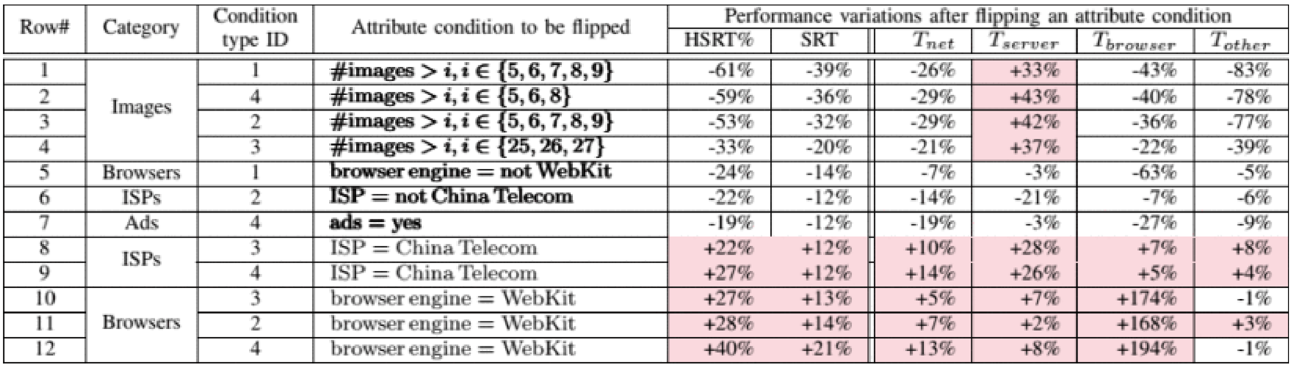


表 III：第一个月（从第 1 天至第 31 天）的普遍状况类型。HSRT 分支属性条件以粗体显示



Condition type ID	Prevalent condition type	Prevalence (days)	HSRT Coverage
1	**#images** $> \boldsymbol{i}, \boldsymbol{i} \in \{5,6,7,8,9\}$ ∧ **browser engine = not WebKit**	21	43%
2	**#images** $> \boldsymbol{i}, \boldsymbol{i} \in \{5,6,7,8,9\}$ ∧ **ISP = not China Telecom** ∧ browser engine = WebKit	15	25%
3	**#images** $> \boldsymbol{i}, \boldsymbol{i} \in \{25,26,27\}$ ∧ ISP = China Telecom ∧ browser engine = WebKit	7	9%
4	**#images** $> \boldsymbol{i}, \boldsymbol{i} \in \{5,6,8\}$ ∧ ISP = China Telecom ∧ browser engine = WebKit ∧ **ads = yes**	6	9%

表 IV：属性条件的影响，按 HSRT%列排序。突出显示大于零的变化



Row#	Category	Condition type ID	Attribute condition to be flipped	Performance variations after flipping an attribute condition					
				HSRT%	SRT	T_{net}	T_{server}	$T_{browser}$	T_{other}
1	Images	1	**#images** $> i, i \in \{5,6,7,8,9\}$	-61%	-39%	-26%	+33%	-43%	-83%
2		4	**#images** $> i, i \in \{5,6,8\}$	-59%	-36%	-29%	+43%	-40%	-78%
3		2	**#images** $> i, i \in \{5,6,7,8,9\}$	-53%	-32%	-29%	+42%	-36%	-77%
4		3	**#images** $> i, i \in \{25,26,27\}$	-33%	-20%	-21%	+37%	-22%	-39%
5	Browsers	1	**browser engine = not WebKit**	-24%	-14%	-7%	-3%	-63%	-5%
6	ISPs	2	**ISP = not China Telecom**	-22%	-12%	-14%	-21%	-7%	-6%
7	Ads	4	**ads = yes**	-19%	-12%	-19%	-3%	-27%	-9%
8	ISPs	3	ISP = China Telecom	+22%	+12%	+10%	+28%	+7%	+8%
9		4	ISP = China Telecom	+27%	+12%	+14%	+26%	+5%	+4%
10	Browsers	3	browser engine = WebKit	+27%	+13%	+5%	+7%	+174%	-1%
11		2	browser engine = WebKit	+28%	+14%	+7%	+2%	+168%	+3%
12		4	browser engine = WebKit	+40%	+21%	+13%	+8%	+194%	-1%

C.

D. Observations by Further Investigation

表 IV 的一些观察和问题，以及结合领域知识进行解释

V. O PTIMIZATION OF H IGH SRT IN P RACTICE