PracExtractor: Extracting Configuration Good Practices from Manuals to Detect Server Misconfigurations

Chengcheng Xiang
University of California, San Diego

Andrew Yoo
University of Illinois Urbana-Champaign

Haochen Huang
University of California, San Diego

Yuanyuan Zhou
University of California, San Diego

Shankar Pasupathy
NetApp Inc.

Abstract

Configuration has become ever so complex and error-prone in today’s server software. To mitigate this problem, software vendors provide user manuals to guide system admins on configuring their systems. Usually, manuals describe not only the meaning of configuration parameters but also good practice recommendations on how to configure certain parameters. Unfortunately, manuals usually also have a large number of pages, which are time-consuming for humans to read and understand. Therefore, system admins often do not refer to manuals but rely on their own guesswork or unreliable sources when setting up systems, which can lead to configuration errors and system failures.

To understand the characteristics of configuration recommendations in user manuals, this paper first collected and studied 261 recommendations from the manuals of six large open-source systems. Our study shows that 60% of the studied recommendations describe specific and checkable specifications instead of merely general guidance. Moreover, almost all (97%) of such specifications have not been checked in the systems’ source code, and 61% of them are not equivalent to the default settings. This implies that additional checking is needed to ensure the recommendations are correctly applied.

Based on our characteristic study, we build a tool called PracExtractor, which employs Natural Language Processing (NLP) techniques to automatically extract configuration recommendations from software manuals, converts them into specifications, and then uses the generated specifications to detect violations in system admins’ configuration settings. We evaluate PracExtractor with twelve widely-deployed software systems, including one large commercial system from a public company. In total, PracExtractor automatically extracts 338 recommendations and generates 173 specifications with reasonable accuracy. With these generated specifications, PracExtractor detects 1423 good practice violations from open-source docker images. To this day, we have reported 325 violations and have got 47 of them confirmed as real configuration issues by admins from different organizations.

1 Introduction

1.1 Motivation

Misconfiguration (error in configuration settings) has become one of the major causes of failures in large-scale cloud and Internet systems, as reported by many system vendors [29,45,64] and service providers [20,26,32,34,37,51]. While various fault tolerance and recovery mechanisms are effective in handling hardware and software failures, they are less effective in handling configuration errors [27,32,37]. In 2017, a configuration error at Level 3, an Internet backbone company, caused a nationwide network outage [22]. On March 13th, 2019, the recent outage in Facebook was also caused by a server configuration error, affecting millions of users [52]. In addition to reliability, configuration errors can also lead to security issues [59]. OWASP reports misconfiguration as one of the top 10 most critical web security risks [38]. In 2017, a configuration error of Amazon S3 storage exposed personal information of 200 million U.S. voters [53].

One of the primary reasons for configuration errors is the ever-increasing configuration complexity, especially with system software [60]. Configuration complexity is partially reflected by the large and almost always increasing number of configuration parameters, as well as their configuration constraints and inter-dependency [31,35,44,47,63], which inevitably increase system admins’ error rates [40,46]. For example, MySQL 8.0 has more than 460 configuration parameters. Similarly, Apache httpd 2.4 has more than 550 parameters. Such a high level of complexity makes system configuration an error-prone task.

While research efforts have been attempted on reducing configuration complexity [55,60], it is still a long journey to fully tame the complexity issue. Today, to assist system admins, software vendors typically release user manuals together with their software. A manual describes in detail the name, usages and sometimes constraints of each configuration parameter. It can be in print as a PDF file or accessed electronically as HTML/XML files, providing good guidance
and reference for system admins to configure and manage server software.

Unfortunately, system manuals are quite large, containing hundreds or even thousands of pages. Table 1 lists the numbers of pages in the manuals of ten software, including one commercial software, COMP-A 1, from a large public company. As the table shows, manuals of MySQL, PostgreSQL, CentOS and Hadoop have 2331-5494 pages. COMP-A has 8283 pages in its technical documentation.

With such a daunting number of manual pages, system admins find manuals hard and time-consuming to refer to and understand. As such, system admins often do not refer to them when configuring systems. Instead, they either rely on their own judgment/guesswork or ask for help from other admins [36]. Previous studies have shown that system admins solved only a small proportion of usage problems (4% to 25%) by referring to manuals [21, 33, 36].

However, manuals still contain useful information and ignoring manuals can lead to configuration errors that cause server downtime and data center outages. Figure 1 gives six real-world configuration errors of commercial and open-source software, in which system admins clearly do not follow good practice recommendations in manuals. The misconfigured parameters in these examples were set to incorrect values, leading to problems of systems’ availability, performance and security. Since these incorrect values are totally legal values (i.e. violating no constraints in source code), they cannot be detected by software’s own checking logic, as well as tools that focus on checking for illegal values [63]. However, in all these cases, the corresponding manuals actually have clearly given recommendations on how to set these parameters. Had these recommendations been followed by system admins, these misconfigurations would have been avoided.

Unfortunately, good practices recommended in manuals or other documents are not fully utilized by system admins to avoid configuration errors mainly due to three reasons:

- Recommended practices are spread out in various parts of manuals and cannot be easily found by system admins due to manuals’ bulkiness and poor navigation [36].
- Many good practice recommendations are not always the same as default settings (more details in §2). A recent study shows that admins tend to go with default settings for more than 80% of configuration parameters, and many configuration errors were caused exactly because admins do not change the default setting [60]. As later shown in our evaluation (cf. Table 12), we also found many (997) cases that system admins just went with bad defaults. Had system admins read the recommendations in the manuals, they could have avoided some of these mistakes.
- As shown in all the examples in Figure 1, good practices recommended in manuals are often soft constraints, which usually are not checked inside software. Thereby, the violations of them cannot be detected by previous tools that were built by either inferring configuration specification from the software’s source code [63] or just directly reusing the source code to check configuration [61].

### 1.2 Our Contributions

This paper studies the research questions on whether it is useful to automatically extract good practice recommendations from manuals and use them to detect system admins’ configuration issues, and if so, how to do it. We first collected and studied 261 recommendations from six large open-source software manuals. Our study shows that 60% of the studied recommendations described specific, checkable specifications instead of just general guidance. In addition, almost all (97%) of the checkable specifications are not checked in source codes, and 61% of them are different from the default settings (reasons and details are discussed in §2).

Based on our characteristic study, we build a tool called PracExtractor, which employs Natural Language Processing (NLP) techniques to automatically extract good practice recommendations from manuals, converts them into specifications, and then uses the generated specifications to detect violations in system admins’ configurations.

We evaluated PracExtractor with manuals of twelve widely-deployed software systems, including one from a commercial company with tens of thousands of customers. Overall, PracExtractor automatically extracts 338 recommendations, with a precision of 86% and a recall of 83%. PracExtractor converts 173 recommendations into specifications with reasonable accuracy. For the six “new” manuals not included in our characteristic study, PracExtractor can achieve a precision of 83% for recommendations and 88% for specifications.

To evaluate the capability of detecting real-world misconfigurations, we run PracExtractor against real-world configurations from top-downloaded container images on DockerHub [24]. PracExtractor detects 1423 violations in 853 images. We reported 325 violations to the image maintainers and got 47 confirmed as real configuration issues, including six issues in images with over 1M downloads and 28 in images with over 1K downloads.

Interestingly, in addition to detecting system admins’ configuration problems, PracExtractor also detects a few incorrect...

---

1We are required to keep the company and the product anonymous.

<table>
<thead>
<tr>
<th>Software</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMP-A</td>
<td>8283</td>
</tr>
<tr>
<td>MySQL</td>
<td>5494</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td>3724</td>
</tr>
<tr>
<td>CentOS</td>
<td>2297</td>
</tr>
<tr>
<td>Hadoop</td>
<td>2331</td>
</tr>
<tr>
<td>Htpd</td>
<td>1009</td>
</tr>
<tr>
<td>HBase</td>
<td>787</td>
</tr>
<tr>
<td>Freebsd</td>
<td>726</td>
</tr>
<tr>
<td>Ubuntu server</td>
<td>413</td>
</tr>
<tr>
<td>Zookeeper</td>
<td>181</td>
</tr>
</tbody>
</table>

Table 1: Number of pages in ten popular software’s manuals.
default settings, three of which have already been confirmed by MySQL and Cassandra developers as real bugs. Incorrect default settings can easily cause configuration errors since system admins are most likely to go with the default [60].

## 2 Characteristic Study

Before we build a tool to extract good practice recommendations from manuals, we first collected and studied 261 real-world recommendations from manuals of six widely-deployed systems listed in Table 2. Our study answers two questions: (1) *Is it useful to extract those recommendations from manuals?* If they are all general advice such as “recommend to set it to a large value”, extracting them is not very helpful since they cannot be used as specifications for automatically checking system admins’ settings. In contrast, if the recommendations are clear specifications such as “recommend to set this to greater than 2000”, extracting them out from manuals can help build checkers to detect violations to them. Additionally, have developers already put in their code to check if system admins follow these recommended practices? If so, there is no need to extract them from manuals. Finally, how often are these recommendations the default settings for the corresponding configuration parameters? If they are not default, why? (2) *How difficult is it to extract good practice recommendations from manuals?* In particular, are manuals structured enough for information extraction?

**Observation 1:** 157 (60%) of the studied good practice recommendations are specific instead of just general advice. We manually studied all recommendations and categorized them based on their contents. If a recommendation is about something that is hard to be checked automatically, it is classified as a “general advice” (e.g. Table 3 last row). Otherwise, it is classified as a “clear specification”, which is further categorized into value, usage, correlation, and property by what is recommended, as explained in the caption of Table 2. An example is given for each category in Table 3.

Table 2 shows the number of recommendations of each category. In total, 157 (60%) of the 261 recommendations describe clear specifications that if extracted can be used for automatically checking system admins’ configuration settings. The remaining 104 recommendations are general advice that is hard to check automatically.

**Observation 2:** 152 (97%) of the specific good practices recommended in manuals are not checked in source code. For each recommendation, we manually examine the source code of each software to see if the recommended practices are checked in source code to warn/inform system admins upon violations. Table 4 shows that only five out of the 157 specific recommendations in manuals are checked in source code.

Listing 1 shows an example where a recommended practice is checked in HBase code. In this case, if the practice is violated by system admins, they will be warned to reexamine the setting of this parameter more carefully.

The goal of our work is exactly to generate more checkings like the HBase case shown in Listing 1, i.e. automatically extract good practice specifications from manuals and build a checker to warn system admins when their settings do not follow the recommended practices.

Violations to good practices may not always be configuration errors. However, as previous work [60, 62] has shown,
many system admins simply rely on guesswork or unreliable sources (e.g. online forums) to configure complex server software. If our checker can give a warning like Listing 1 when the settings do not follow practices, system admins can at least have a chance to reexamine the settings more carefully.

**Observation 3:** 96 (61%) of the specific good practice recommendations are not equivalent to default settings. It is conceivable that some recommended practices might be the default settings (after all, the vendor recommends them). If this is the case, there is no need to extract recommended practices from manuals. System admins simply just go with default if they do not know how to set it better.

However, as shown in Table 5, only 61 (39%) good practices are equivalent to the default settings. For the majority (61%) cases, recommendations are not the same as default due to several reasons, including (a) 30 recommend multiple different values, e.g. a range or a set of values. In real settings, they may need to be modified to accommodate different situations, so it is worthwhile for sysadmins to double-check if the settings follow recommendations.; (b) 30 recommend some settings based on some conditions, e.g. “Enable A along with B”; (c) 21 recommendations are on command line options that have no default values; (d) 9 recommend relative values, such as “25% system RAM size”; (e) 6 cases have no clear reason why the default is different. They may be potential bugs and we have one of them confirmed as a bug by developers.

**Observation 4:** The six studied manuals are organized in a similar structure. As shown in Table 6, the six manuals are either in HTML or XML format and parameters in them are described in a similar structure:

- Each parameter is described in one separate section.
- Parameter names are often used as the section headings.
- There is some meta-info of the parameter described in the format of <key>:<value>, such as “Type:string”.
- Most information related to each parameter is described in one or several paragraphs of free texts.

The per-parameter section structure makes it possible to relate each parameter name and its description by parsing the section structure. In addition, data types and default values can be used to identify parameter values in plain text descriptions which is necessary for generating specifications.

### 3 Design and Implementation

We design and implement PracExtractor to automatically extract recommendations from manuals, convert them into specifications and then uses them to detect violations. PracExcel-
Table 6: Format and structure of manuals regarding how they describe configuration parameters. “Parameter section”—a separate section describes each individual parameter, “Data type”/“Default value”—the format they are described in, including table and KV (<key>:<value>).

<table>
<thead>
<tr>
<th>Software</th>
<th>Manual format</th>
<th>Parameter section?</th>
<th>Data type?</th>
<th>Default value?</th>
</tr>
</thead>
<tbody>
<tr>
<td>MySQL</td>
<td>html</td>
<td>Yes</td>
<td>Table</td>
<td>Table</td>
</tr>
<tr>
<td>Httpd</td>
<td>xml</td>
<td>Yes</td>
<td>No</td>
<td>Table</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td>html</td>
<td>Yes</td>
<td>KV</td>
<td>Text</td>
</tr>
<tr>
<td>HDFS</td>
<td>xml</td>
<td>Yes</td>
<td>No</td>
<td>Table</td>
</tr>
<tr>
<td>HBase</td>
<td>html</td>
<td>Yes</td>
<td>No</td>
<td>KV</td>
</tr>
<tr>
<td>Spark</td>
<td>html</td>
<td>Yes</td>
<td>No</td>
<td>Table</td>
</tr>
</tbody>
</table>

Table 7: 10 sample keywords (words and bigrams) collected by PracExtractor from 261 studied recommendations and how many recommendations each covers.

<table>
<thead>
<tr>
<th>Word</th>
<th>Covered sentences</th>
<th>Bigram</th>
<th>Covered sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>recommend</td>
<td>74</td>
<td>be recommended</td>
<td>34</td>
</tr>
<tr>
<td>well</td>
<td>26</td>
<td>should only</td>
<td>20</td>
</tr>
<tr>
<td>good</td>
<td>26</td>
<td>may want</td>
<td>13</td>
</tr>
<tr>
<td>appropriate</td>
<td>21</td>
<td>good idea</td>
<td>7</td>
</tr>
<tr>
<td>want</td>
<td>17</td>
<td>with caution</td>
<td>7</td>
</tr>
</tbody>
</table>

3.2 Recommendation Sentences Extraction

Most sentences in manuals do not contain recommendations. For the twelve evaluated software manuals, 696–25510 sentences are extracted from parameter sections, but only 0.4%–2.7% of them contain recommendations. To extract these small percentage of recommendations, PracExtractor performs two steps filtering:

Keyword-based Filtering Following the intuition that recommendations are usually described with certain keywords (e.g. “recommend”, “suggest”), PracExtractor extracts candidate recommendations with keyword filtering. To find out which words/phrases should be used as the keywords, PracExtractor first breaks the studied 261 recommendation sentences into individual words and bigrams (two consecutive words) and uses them as the candidate keywords $T$. PracExtractor then uses inverse document frequency (IDF) to rank the candidate keywords. IDF reflects how frequently a term $t$ (word/bigram in our case) occurs in a set of sentences set $S$, as:

$$IDF(t, S) = \log \frac{|S|}{\sum_{s \in S} 1_{t \in s}}.$$

PracExtractor calculates $IDF(t, R)$ for the studied recommendations $R$ and $IDF(t, S)$ for all the manual sentences $S$. PracExtractor ranks $T$ based on $IDF(t, R)$ and $IDF(t, S)$ and get the smallest 100 and 300 terms separately as $T_R$ and $T_S$. PracExtractor uses $T_R \cap T_S$ as the final keywords. The intuition behinds this is to find the words that are important in recommendations but not normal sentences. In Table 7, the sample keywords show that PracExtractor has effectively found keywords related to recommendations.

Syntactic-Pattern-based Filtering Using keyword filtering alone is not enough. After keyword-based filtering, only 7.3% of the remaining sentences are recommendations. Many sentences with the recommendation-related keywords are not true recommendations. Figure 2 (a) and (b) gives examples that the same keywords can be contained both in recommendation and non-recommendation sentences.

The key difference between these recommendations and non-recommendations in Figure 2 is their syntactic patterns.
Besides a keyword, the recommendations also contain a setting phrase, a noun/verb phrase describing what setting is recommended. Between such setting phrases and keywords, there are certain syntactic relations (patterns), which do not exist in non-recommendation sentences. PracExtractor leverages the syntactic-patterns to do fine-grained filtering.

PracExtractor first adopts the universal dependency (UD) tree [23] to represent a sentence’s syntactic structure. A UD tree $T = (V, E)$ consists of vertices $V$ and edges $E$, where $v \in V$ is labeled with a word’s part of speech (POS) and $e \in E$ represents the syntactic dependency between two words (cf. Figure 2). Let $T' = (V', E')$ be an undirected correspondence of $T$, the syntactic pattern between a keyword and a setting phrase can be represented with an undirected path $p = (v_0, e_{v_0,v_1}^i, v_1, \ldots, v_n)$, where $v_0$ is the keyword, $v_n$ is the setting phrase, and $e_{v_{i-1},v_i}^i \in E'$ for $i \in [1, n]$.

With the UD representation, PracExtractor mines the unique patterns for recommendations from the studied 261 recommendations $S_{rec}$ and a set of non-recommendation samples $S_{not\_rec}$ that contains the keywords. For each sentence $s$, PracExtractor builds $T'_s = (V_s, E'_s)$ and extracts all paths

$$\rho_s = \{(v_0, e'^{v_0,v_1}_0, v_1, \ldots, v_n) :$$

$$v_0 \in \text{KEYWORDS} \land \forall i \in [1, n] \quad e'^{v_i,v_{i+1}}_i \in E'_s$$

$$\land \forall i \in [0, n] \quad v_i \in V_s \land v_n \in \text{SETTINGPHRASES}\},$$

that starts from each keyword and ending at each setting phrase. The keywords are from the last step and the setting phrases are labeled by human inspectors. PracExtractor extracts all such paths from all recommendations and non-recommendation samples, denoted as $P_{rec}$ and $P_{not\_rec}$. PracExtractor then extracts patterns $P_{pattern}$ with Algorithm 1.

With the identified syntactic patterns $P_{pattern}$, PracExtractor classifies a new sentence $s'$ into a recommendation or non-recommendation. PracExtractor traverse $P_{pattern}$ and check if any pattern matched with $s'$. If at least one pattern matched then $s'$ is classified as a recommendation otherwise non-recommendation. Such a matched pattern also labels the setting phrase in $s'$ at the pattern’s end (cf. Figure 2). The setting phrase will be used in specification generation (cf. §3.3).

### 3.3 Specification Generation

In §3.2, PracExtractor identifies recommendation sentences and the setting phrases within it. PracExtractor then converts the setting phrases into checkable, formal specifications. Table 8 gives three example recommendations and the corresponding specifications. In general, PracExtractor can generate four types of specifications, including value, correlation, usage and property, as shown in Table 9.

A naive way to generate specifications is to match setting phrases with predefined regular expressions and convert them accordingly. This can transform simple phrases with numbers (e.g. “less than 8”), but cannot convert more complex phrases (e.g. phrases with enum or parameter names). For instance, a phrase could be “set to chain”, where “chain” is an enum value in Httpd. Such software-specific words can hardly be predefined in regular expressions and so cannot be matched.
which parameter a sentence is associated with. For a recom-
mendation sentence, PracExtractor identifies this sentence is for

**Table 8: Examples of specifications generated by PracExtractor.**
Setting phrases are marked with rectangles.

<table>
<thead>
<tr>
<th>Category</th>
<th>Specification</th>
<th>Description</th>
<th>Patterns</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>value/correlation</td>
<td>p == v</td>
<td>v_value</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p &lt; v</td>
<td>less_syn _more_syn than v_value</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p e {v, v}</td>
<td>between_syn v_value to v_value</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p e {v, v}</td>
<td>v_value or v_value = # _value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>correlation</td>
<td>with (p, p')</td>
<td>along_syn with p_value</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>prefer (p, p')</td>
<td>prefer_syn p_value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>usage</td>
<td>use (p)</td>
<td>used_syn</td>
<td></td>
<td></td>
</tr>
<tr>
<td>property</td>
<td>format (p, f)</td>
<td>f_format</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 9: Category of specifications PracExtractor generates and example of patterns for each specification.** “<value>” is defined as “<bool><num><unit><enum><parameter>” from Table 10. “less\_syn” means the synonyms of “less”.

PracExtractor addresses this issue in three steps. First, given a recommendation sentence, PracExtractor identifies which parameter the sentence is associated with. Then, PracExtractor uses the parameter’s meta-info (e.g. type and default value) extracted before (cf. §3.1) to identify setting entities, such as values and formats. Third, PracExtractor matches the identified setting entities with predefined semantic patterns to generate specifications.

**Identify Parameter Names** PracExtractor first identifies which parameter a sentence is associated with. For a recommendation sentence \( s \) in a paragraph \( p \) related to parameter \( X \), there are four possible cases: 1) Only \( X \) is mentioned in \( s \). PracExtractor determines this sentence is for \( X \); 2) Another parameter \( Y \) is mentioned in \( s \). PracExtractor checks if \( Y \) is a subject or object to a verb like “set” or “specify” and determines the sentence is for \( Y \) if it is the case; 3) No parameter is mentioned in \( s \). PracExtractor further searches previous sentences in paragraph \( p \); 4) No parameter is mentioned in \( p \), PracExtractor determines the sentence is for \( X \).

**Identify Settings Entities** Give the identified setting phrase and associated parameter of a sentence, PracExtractor then recognizes setting entities (e.g. values and formats) from the setting phrase based on the associated parameter’s type. Table 10 shows the seven types of setting entities that PracExtractor can identify. For different types of setting entities, PracExtractor identifies them with different syntax:

- For basic types, including <bool> and <num>, PracExtractor identifies them with regular expressions. For <num>, PracExtractor also identifies common <unit> (e.g. “GB”, “byte”) along with it.
- For <enum>, PracExtractor takes advantage of parameters’ type-info to identify it. The type-info of an <enum> parameter does not only indicate it is <enum> but also indicates the valid values that can be set to the parameter. PracExtractor has parsed these in the parsing step (cf. 3.1) and now searches a setting phrase for the valid values.
- <parameter> is for the case that the setting phrase describes current parameter with respect to another parameter, such as “set A to be larger than B”. PracExtractor identifies this by searching for valid parameter names in the setting phrase.
- For <format>, PracExtractor identifies common formats (e.g. “email address”, “absolute path”) based on word matching. PracExtractor allows users to provide new words to extend the identifiable formats.
- All other words/phrases are identified as <string>. PracExtractor further handles two kinds of strings that have special meaning. First, some of them use “this value” or “this” to refer to a value mentioned in previous sentences. PracExtractor recognizes such references and identifies the actual value from previous sentences. Second, some may use “default” to refer to parameters’ default value. In this case, PracExtractor uses the corresponding default value extracted from the parsing step (cf. §3.1) as the recommended setting.

**Generate Specification** PracExtractor generates checkable formal specifications by matching the setting phrases with predefined semantic patterns. Table 9 lists the types of speci-
ifications that can be generated by PracExtractor, each with an example pattern. Before blindly matching a setting phrase with patterns, PracExtractor first considers the associated parameter’s data type. For example, if the data type is <bool> or <enum>, there is no need to match semantic patterns like “less than <value>”, “between <value> and <value>”, or anything that is for <num>. Second, synonyms are also considered for these patterns. For instance “less<syn>” can also match with “lower than” and “smaller than”, etc.

Detect Negation PracExtractor also detects negation in a recommendation sentence and negates a specification when necessary. Two different types of negations are handled by PracExtractor. First, PracExtractor finds direct negation words, such as “not” (or abbreviation “n’t”) “none” and “never”, in sentences. In addition, PracExtractor also detects indirect negation words and phrases, such as “avoid”, “with caution”, “rarely”, and “seldom”, etc. to identify the negation.

3.4 Violation Detection from Configurations

PracExtractor parses a configuration file and turns the settings into key-value pairs of (parameter name, value). Although the formats of configuration files for different software can vary depending on the software implementation, most of them have similar formats. The configuration files of the twelve popular systems in our evaluation, including MySQL, Httpd, HBase, HDFS, Spark, Squid and even the commercial system, all follow two common and simple formats: key-value pairs with separator like ‘=’ or ‘:’ or XML format.

Then PracExtractor checks the parsed parameter values against extracted specifications and generates warning messages if it detects violations. PracExtractor uses the original sentences from manuals as the warning messages. An example warning could be “Setting dfs.safemode.replication.min lower than dfs.replication.min is not recommended and is dangerous for production setups”. This warning can remind sysadmins to double-check the configurations to avoid potential mistakes, just like the one shown in Figure 1 and 3.

Most of the violations detected by PracExtractor cannot be detected by previous works. Previous works [44, 61, 63] mainly use configuration checking/usage logic in source code to detect misconfigurations. However, most recommendations (97% as in Table 4) are not checked in source code, so violations to them cannot be detected by these works.

4 Experimental Evaluation

As shown in Table 11, we evaluate PracExtractor on twelve large software systems including eleven popular open-source server systems and one commercial systems (COMP-A) from a public company that serves many enterprise customers. These systems’ manuals have 543 to 8283 pages. These manuals include not only the six manuals in our characteristic study, but also six new manuals that are not studied before. We evaluate both the precision and recall of PracExtractor by comparing its results with recommendations identified by human inspectors. In our evaluation, two human inspectors manually and independently examined each manual to identify recommendations.

For violation detection, we evaluate PracExtractor with real-world settings from top-ranked container images on DockerHub (200 images for each open-source systems). We run PracExtractor against configuration files in these docker images to detect violations to the specifications extracted from manuals. The evaluated images include both Linux-based and Windows-based one. However, as PracExtractor currently does not support platform-specific checking, our evaluation does not include checking platform-related specifications. In total, only 4 specifications are platform-related.

Recommendation and Specification Extracted As shown in Table 11, PracExtractor extracts a total of 338 good practice recommendations (including specific and general advice) from manuals and automatically converts 173 of them into formal specifications that can be used to check system admins’ configuration settings. Among all software, Httpd has the most number (81) of recommendations extracted as well as the most number (31) of specifications generated.

Results with the six “new” manuals excluded from our study PracExtractor works reasonably well with the six “new” software manuals that are not included in our characteristic study. PracExtractor extracts 117 recommendations and 59 specifications from these manuals. For example, it extracts 35 recommendations and 22 specifications for the commercial software, COMP-A. The precision and recall are only slightly lower than the one for the six studied manuals, but are still reasonably good (with a 0.83 precision and 0.80 recall for recommendations, and 0.88 precision and 0.66 recall for specifications).

Violations Detected PracExtractor detects in total 1423 practice violations from 853 unique images. We manually validated all the violations. We reported 325 (that are maintained on GitHub) of them to their maintainers and have got 47 confirmed as real configuration issues, including six in images with >1M downloads and 28 in images with >1K downloads.

Table 12 shows a breakdown of all detected violations. 426 are “wrong change”, namely a parameter is explicitly changed to a non-recommended value by system admins. 997 violations are “wrong default”, namely a parameter has a non-recommended default value but is not changed. This also matches with the finding from a previous real-world misconfiguration study [60] that many configurations are left as default. In this case, system admins are afraid of changing default settings, even though the default is not recommended or is simply a placeholder value, which needs system admins to explicitly change to fit their own system environments.
Table 11: Numbers of recommendations extracted and specifications generated by PracExtractor and corresponding accuracy (precision and recall). Recommendations consist of both general advice and specific ones that can be converted into specifications. “studied” refers to the software included in our characteristic study (first 6 rows), while “new” refers to other software not included in our study (last 6 rows).

Table 12: Detected good practice violations in container images from Dockerhub. We reported 325 violations to the image owners, and 47 of them have been confirmed as real configuration errors. Three wrong defaults are also confirmed by MySQL and Cassandra.

Figure 3 gives three examples of real-world violations that are detected by PracExtractor from popular container images on DockerHub. They have been confirmed by the image owners as real configuration errors or by the software developers as real bugs. Here are the root causes:

(a) HDFS manual recommends to leave the parameter as true to avoid registration of excluded hostnames. However, it is ignored and violated in 21 images, which can cause security issues. We reported them and so far two of them have been confirmed [11, 12].

(b) Cassandra manual does not recommend to enable the experimental feature as it may cause potential failure. However, the default setting enables it, which is a bug, and 199 images just keep the default. The bug has been confirmed and fixed by Cassandra in its new version [1].

(c) The default setting for this parameter in MySQL is much larger than the recommended value in the manual, and the default value (set by MySQL developers) is actually incorrect. We report it to MySQL official Bugzilla and it has been confirmed as a bug [16].

Indeed, not all the 1423 violations are configuration errors or bugs. However, as discussed before in the real-world example from HBase (cf. §2 Listing 1) that explicitly performs such checks in its source code, when such violations are warned to system admins, they at least get a chance to reexamine and reconsider the settings more carefully.

Maintainers’ Feedback on Violations We reported 325 to the image maintainers and so far have got 47 violations confirmed that they need to be changed. We list three example confirmations in Table 13 (row 1, 2, 3). We took a further look into the impact of these confirmed violations: 11 cause security vulnerabilities, 31 cause unreliable services, 2 cause performance issues and 3 cause database inconsistency.

We also got 46 other feedbacks that the maintainers hesitated to fix the violations. They either think it is the upstream vendors’ responsibility to handle the issues (Table 13 row 4, 5) or are aware of the limitations but make the settings for particular environments (Table 13 row 6).

Lines of Customized Code for Each Manual Table 14 shows the lines of Python code (LOC) of the format spec part of Manuals.

## Table 11: Numbers of recommendations extracted and specifications generated by PracExtractor and corresponding accuracy (precision and recall). Recommendations consist of both general advice and specific ones that can be converted into specifications. “studied” refers to the software included in our characteristic study (first 6 rows), while “new” refers to other software not included in our study (last 6 rows).

<table>
<thead>
<tr>
<th>Software</th>
<th>Category</th>
<th>Update Time of Manuals</th>
<th># Recommendations</th>
<th># Specifications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>total</td>
<td>extracted</td>
</tr>
<tr>
<td>MySQL</td>
<td>database</td>
<td>Aug. 2019</td>
<td>78</td>
<td>61</td>
</tr>
<tr>
<td>Httpd</td>
<td>web server</td>
<td>Aug. 2019</td>
<td>92</td>
<td>81</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td>database</td>
<td>Aug. 2019</td>
<td>49</td>
<td>38</td>
</tr>
<tr>
<td>HDFS</td>
<td>distributed storage</td>
<td>Aug. 2019</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>HBase</td>
<td>distributed storage</td>
<td>Aug. 2019</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Spark</td>
<td>distributed computing</td>
<td>Aug. 2019</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>COMP-A</td>
<td>commercial storage</td>
<td>May 2019</td>
<td>49</td>
<td>35</td>
</tr>
<tr>
<td>Nginx</td>
<td>proxy</td>
<td>Jul. 2019</td>
<td>26</td>
<td>24</td>
</tr>
<tr>
<td>Flink</td>
<td>stream processing</td>
<td>Aug. 2019</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Squid</td>
<td>proxy</td>
<td>Feb. 2019</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>Mapred</td>
<td>distributed computing</td>
<td>Aug. 2019</td>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>Cassandra</td>
<td>distributed storage</td>
<td>Aug. 2019</td>
<td>15</td>
<td>14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Studied</th>
<th>New</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Studied 261</td>
<td>0.89 0.85 157</td>
</tr>
<tr>
<td></td>
<td>New 147</td>
<td>0.83 0.80 90</td>
</tr>
<tr>
<td></td>
<td>Overall 408</td>
<td>0.86 0.83 247</td>
</tr>
</tbody>
</table>

## Table 12: Detected good practice violations in container images from Dockerhub. We reported 325 violations to the image owners, and 47 of them have been confirmed as real configuration errors. Three wrong defaults are also confirmed by MySQL and Cassandra.

<table>
<thead>
<tr>
<th>Software</th>
<th>Wrong change</th>
<th>Wrong default</th>
</tr>
</thead>
<tbody>
<tr>
<td>MySQL</td>
<td>20 200</td>
<td></td>
</tr>
<tr>
<td>Httpd</td>
<td>338 200 8</td>
<td>0</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td>8 0</td>
<td>199</td>
</tr>
<tr>
<td>HDFS</td>
<td>21 0</td>
<td></td>
</tr>
<tr>
<td>HBase</td>
<td>0 199</td>
<td></td>
</tr>
<tr>
<td>Spark</td>
<td>10 0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>426</td>
<td>997</td>
</tr>
</tbody>
</table>
**Accuracy — False Negatives and False Positives**

Table 13: Example of positive and less-positive (in gray background) feedback. Due to page limit, we list three for each.

<table>
<thead>
<tr>
<th>Software</th>
<th>Manual LOC pages</th>
<th>LOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MySQL</td>
<td>5494</td>
<td>73</td>
</tr>
<tr>
<td>Httpd</td>
<td>1009</td>
<td>10</td>
</tr>
<tr>
<td>PostgreSQL</td>
<td>3724</td>
<td>24</td>
</tr>
<tr>
<td>HDFS</td>
<td>1031</td>
<td>12</td>
</tr>
<tr>
<td>HBase</td>
<td>787</td>
<td>14</td>
</tr>
<tr>
<td>Spark</td>
<td>599</td>
<td>6</td>
</tr>
<tr>
<td>Software</td>
<td>Manual LOC pages</td>
<td>LOC</td>
</tr>
<tr>
<td>COMP-A</td>
<td>8283</td>
<td>18</td>
</tr>
<tr>
<td>Nginx</td>
<td>543</td>
<td>24</td>
</tr>
<tr>
<td>Flink</td>
<td>6152</td>
<td>6</td>
</tr>
<tr>
<td>Squid</td>
<td>1391</td>
<td>10</td>
</tr>
<tr>
<td>Mapred</td>
<td>1318</td>
<td>12</td>
</tr>
<tr>
<td>Cassandra</td>
<td>913</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 14: Lines of code (LOC) of format specs for the twelve evaluated manuals.

**Impacts of False Positives** The false positives will not cause serious impacts as they can be recognized easily. Table 16 shows three cases of false recommendations that PracExtractor extracted from the evaluated manuals. They are descriptions related to the parameters but just are not recommendations. For example, “there may be circumstances where it is desirable for a configuration section’s authorization to be combined with that of its predecessor” is a false positive, which describes a parameter usage but gives no recommendation. Since PracExtractor includes such an original sentence from manuals in a warning message, sysadmins who read it can easily realize that it is not a recommendation and will not be misguided.

**Evaluation with Existing Misconfiguration Dataset** We also evaluate PracExtractor with existing configuration issues. As there is no existing dataset for good practice violations, we use a dataset [2] for general configuration issues used in previous works [60, 61]. This dataset contains configuration issues from various online forums and mailing lists. We use the issues categorized as “error” to evaluate PracExtractor. Out of the 63 evaluated configuration errors, PracExtractor can detect 7 (10%) of them. We validated that these detected errors cannot be detected by previous works [35, 61] as there violate no constraint in source code while previous works use code constraints to detect errors. For the undetected errors, we found that manuals do not provide recommendations for
Table 15: Root causes for PracExtractor’s False Negative and False Positive. R1 — descriptive recommendations that are not identified, such as “it is recommended to not configure a ticket key file”. It is hard to automatically infer that this refers to not using SSLSessionTicketKeyFile. R2 — unknown string values that are not identified. For example, in “it’s recommended the username ‘anonymous’ is in allowed userIDs”, it is hard to recognize the common word “anonymous” as a value. R3 — sentences that are not covered by our syntactic patterns. R4 — texts that are incorrectly identified as parameter values. R5 — non-recommendation sentences that are mismatched with syntactic patterns of recommendations.

<table>
<thead>
<tr>
<th>Software</th>
<th>False Negatives</th>
<th>Software</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Httpd</td>
<td>10 11 0</td>
<td>Httpd</td>
<td>4 4</td>
</tr>
<tr>
<td>COMP-A</td>
<td>8 3 4</td>
<td>Nginx</td>
<td>4 0</td>
</tr>
<tr>
<td>Mapred</td>
<td>1 2 3</td>
<td>Squid</td>
<td>2 0</td>
</tr>
</tbody>
</table>

Table 16: Example of false positive recommendations PracExtractor extracted. They can be easily recognized by system admins and will not misguide them to make wrong changes.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>False Positive Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>adaptive_hash_index</td>
<td>It may be desirable to dynamically enable or disable adaptive hash indexing to improve query performance.</td>
</tr>
<tr>
<td>AuthMerging</td>
<td>There may be circumstances where it is desirable for a configuration section’s authorization to be combined with that of its predecessor.</td>
</tr>
<tr>
<td>LimitRequest</td>
<td>When name-based virtual hosting is used, the value for this directive is taken from the default (first-listed) virtual host best matching the current IP address and port combination.</td>
</tr>
</tbody>
</table>

5 Discussion

Generality. PracExtractor is reasonably general for manuals from different software. As our evaluation on six new manuals shows, PracExtractor can identify most (80%) recommendations from these manuals with a precision of 83%.

Human Effort. PracExtractor can easily be extended with new format specs to accommodate future manuals. In our evaluation, the specs for the six new software manuals are all less than 30 lines (cf. Table 14) and were written by a first-year graduate student, each in 0.5-2 hours.

Accuracy of Manuals. Another concern is whether manuals themselves deliver accurate information for PracExtractor to extract. After all, manuals can be outdated and can have mistakes made by the writers. In our work, we considered these factors. First, we validated the last update time of our evaluated manuals. As shown in Table 11, all twelve manuals are updated recently this year. Second, we compared the specifications extracted from manuals against source code and reported all differences to developers to check. If either is wrong/obsolete, it may introduce problems. Interestingly, in three cases, developers from MySQL and Cassandra confirmed that the source code is wrong (cf. Figure 3).

6 Related Work

Misconfiguration detection and troubleshooting. Many works have been done on detecting [25, 28, 30, 61, 66, 68] and troubleshooting [17–19, 42, 43, 54, 56, 57, 65, 69] configuration errors. Almost all of these works detect configuration errors by either checking against (a) patterns mined from tons of configuration files, or (b) constraints inferred from source code. Our work is complementary to these approaches. We extract recommendations from vendor-provided manuals as specifications, use them to detect violations, and warn system admins to reexamine the violations. Each of the approaches has its own strengths and weaknesses. Below, we compare our approach with each previous one respectively.

Xu [61,63], Rabkin [44], and Nadi [35] propose approaches to extract configuration constraints from source code using static analysis. While it can infer simple constraints such as parameter types, range and some simple dependencies, it is less effective in checking against more complex constraints, especially configuration settings that are legitimate but may not be good or optimal. In our work, we focus on good practices recommended by vendors. If a system admin configures a parameter with a valid setting but does not achieve the intended goal (e.g. performance, reliability or security goal), previous works that focus on detecting invalid configurations will not report any problem. In comparison, our PracExtractor can still warn him/her about the setting if it does not follow the good practices PracExtractor extracts from manuals.

Encore [68], PeerPressure [54] and Santolucito et al’s work [47,48] propose to extract configuration constraints and good practices from existing settings. These approaches assume that a large number of independent configuration samples are available for learning and the correct configurations are the common ones. This assumption can be true for some systems where configuration settings can be collected from users/customers back to vendors. However, for many enterprise software such as database or storage systems that are mainly deployed in enterprises (e.g. financial companies and government), each system’s configuration settings are confidential information and cannot be shared back with vendors. As such, these approaches are less applicable. In comparison, PracExtractor is applicable to such scenarios because it automatically extracts good practice recommendations that are already written in vendor-provided manuals, and the check-
ers generated by PracExtractor can be shipped to enterprise customers to check their configuration settings.

In addition to the above two approaches, another closely related work is ConfSeer [41], which takes a user’s configuration of one or multiple parameters to search against the vendor’s Knowledge Base (KB) articles and identifies those highly relevant ones so that users can read those KB articles to self-diagnose and self-correct misconfiguration (with no need to call customer support). While this work also uses NLP techniques, their goal is to narrow down the match so that users do not need to read hundreds of KB articles. For a given configuration parameter setting from a user, ConfSeer returns a ranked list of KB articles for the user to explore, in a way similar to a search engine like Google that tries to return the most relevant web pages to a user’s query. In other words, ConfSeer is an improved Google search engine for configuration-related KB articles. In comparison, our goal is to extract configuration recommendations from manuals and convert them into formal and checkable specifications. Our checker can be shipped to the customer sites to automatically detect violations and warn system admins to reexamine their settings to proactively avoid problems instead of waiting for postmortem troubleshooting.

Inferring specification from text. Some past works also aim to infer specifications from program-related texts, including program comments [49, 50], API documents [39, 67, 70, 71], and man pages [58]. Our work differs from previous works both on purposes and techniques. First, instead of helping developers find bugs in source code as in [49, 50, 70], our work aims to help system admins detect misconfigurations in their system settings. Secondly, comments and API documents have relatively uniform structures, which makes it easy to extract information like function names and variable types. In comparison, manuals are much less structured. The only structure is that each parameter has its own section/chapter. Inside a section, it is mostly free text. Thirdly, PracExtractor extracts much more complex constraints than previous work on man pages [58]. DASE [58] uses regular expression to extract valid options from man pages. In comparison, PracExtractor can extract option value, correlation and property from software manuals.

7 Conclusions and Future Work

This paper focused on the usefulness and feasibility of extracting good practice recommendations from software manuals to detect configuration problems. Specifically, we first conducted a characteristic study on 261 recommendations from six large open source software manuals. Based on the observations learned, we designed and implemented a tool, called PracExtractor, that can extract 338 recommendations and generate 173 specifications with reasonable accuracy from twelve large software manuals, including one for a large commercial software system. Additionally, with the generated specifications, PracExtractor have detected 1423 violations from 853 container images on DockerHub. We reported 325 of them and so far have got 47 confirmed as real configuration issues by the image maintainers from different organizations.

Interestingly, in addition to detecting system admins’ configuration problems, PracExtractor can also help detect incorrect default settings for configuration parameters. When a default setting differs from a recommendation in the manual, it may indicate that the default setting is wrong. Incorrect default settings can easily cause configuration errors because system admins are most likely to go with default [60]. In our experiments, we did discover a few such software bugs, and three of them have already been confirmed respectively by MySQL and Cassandra.

PracExtractor is far from perfect. First, there is still much space to further improve its accuracy based on our analysis of false positives and false negatives (cf. Table 15). Further semantic analysis of parameter descriptions can improve the identification accuracy of parameter type, name and value. Second, PracExtractor currently cannot extract specifications with descriptive conditions, such as “set A with a large workload”. This may possibly be handled by further incorporating domain knowledge of common descriptions and sub-clause analysis techniques.

Several other directions are also valuable to explore in the future. First, from our experience, we found that a uniform structure (e.g. per-parameter section) and consistent word usage (e.g. recommend) benefit extracting recommendations a lot. Therefore, it would be interesting to explore how manuals may be restructured so more information can be automatically extracted. Second, PracExtractor may be potentially used to detect documentation drift — manuals are not updated along with source code. By combining PracExtractor with source code analysis tools, it is possible to compare the configurations described in manuals and used in source code. Third, while PracExtractor focuses on analyzing user manuals, the approach may be applicable to extract good practices from other text-based documents such as knowledge-base (KB) articles, which are used by support engineers to troubleshoot customer issues.

Acknowledgments

We greatly appreciate the anonymous reviewers and our shepherd, Scott D. Stoller, for their insightful comments and feedback. We thank the Opera group, the Systems and Networking group at UCSD as well as Anita Jinda, Sam Lin and Jianmo Ni for useful discussions and paper proofreading. We thank Tianyin Xu for sharing a configuration issue dataset with us. This work is supported in part by NSF grants (CNS-1814388, CNS-1526966) and the Qualcomm Chair Endowment.
References

[1] Cassandra release note. https://gitbox.apache.org/repos/asf?p=cassandra.git;a=blob;f=NEWS.txt;h=ead2f0ac3d8d93c1ad87a3b944c0c72345257c1;hb=HEAD.


