Crawlzilla - A Toolkit for Deploying Cluster Search Engine Quickly and Easily

Shun-Fa Yang, Wei-Yu Chen, Wen-Chieh Kuo
National Center for High-Performance Computing
Free Software Lab
Taichung, Taiwan
Email: {shunfa, waue, rock}@nchc.org.tw

***Abstract*—Nutch is one of the most well-know and best search engine project for crawling enterprise or personal internal web sites, but many system administrators encounter difficulties to setup and use due to the complicated operation process. In this paper, we present Crawlzilla, an open source search engine tool built on top of Hadoop and Nutch.**

Crawlzilla integrates related useful packages to reduce installation and setup steps, assists system administrators to deploy their own private search engine within the intra website quickly, and also supplies cluster feature to build distributed search engine environment. In addition, it also provides two friendly interfaces for system administrators. The one is used to manage system environment operated on terminal window, the other interface based on web page help system administrators or users for creating their own search engines.

Keywords-Search Engine, Nutch, Hadoop, Java Open Source

I. INTRODUCTION

Nowadays, the web pages are increasing very fast. How to help system administrators to find out the correct information is the fundamental goal for search engine. Therefore, search engine becoming more and more necessary and popular in surfing the Internet. However, these famous search engines, such as Google or Yahoo, are only working for public internet but confidential website. Either, we cannot customize these search engines for special purpose. Based on these factors, open source search engine is very adapted for intra website or customized usage. Nutch is one of the most famous and best open-source search engine project, adapted to personal and business usage. However, using Nutch is not easy for system administrators, especially for those system administrators who are not familiar with the tedious setup and complicated operation process. System administrators encounter more obstacles, such as cluster setting and detail configuration. Therefore, we develop a search engine tool, Crawlzilla, for better system administrators experience. This system is integrated Nutch with related packages and provides easy installation and operation. Crawlzilla can automatically distribute jobs to each computing slave nodes by Hadoop Map-Reduce framework. Besides, Crawlzilla provides two main interfaces, both are able to be operated from remote site. The interface for operation based on web page helps system administrators for creating their own search engines. The other one is used to manage entire cluster environment including the Namenode, Datanode, Jobtracker, Tasktracker and web service.

In the following sections, we provide more details about Crawlzilla architecture and capabilities. Section II describes the related components Nutch and Hadoop. Section IV details Crawlzilla design and architecture. Section V describes the implement of Crawlzilla. Future work is concluded in Section VII.

II. BACKGROUND

In this section, we introduce the operation method then focus on the relation between Hadoop and Nutch.

A. Nutch

Nutch [1] is an open-source software, which contains modules for crawling, indexing and searching. System administrators are required to provide a file containing seed urls.

Nutch crawls predefined number of web pages starting from the given seed urls. As shown in Figure 1, first injector injects seed urls into crawl database as unfetched urls. Crawl database contains <url, crawl-data> as <key, value pair>, where crawl-data contains necessary information of url. The information is whether it is fetched, unfetched or linked, last fetch time, signature, etc. Now generate-fetch-parse-update cycle runs depth times, which is defined parameter by system administrators. In each cycle a new segment is generated, which contains all the required data. Generator module selects unfetched urls from crawl database and puts in fetch list for newly segment. The generator task selects best score urls to processes <url, crawl-data> records from crawl database and parses as (inlink score)-<url, crawl-data> pair to the run-time system. Then, the <url, crawl-data> pairs are instead of (inlink score)-<url, crawl-data> pairs. Fetcher module fetches all the urls from the fetch list and store the web pages in content database. Parser module analyzes web pages and generates url-parsed databases. Once the crawling is completed, invertlink module inverts all links using parsed data to get anchor text. Indexer module generates the IndexDB with anchor text and parsed text.

 ISBN: 978-1-61208-135-9

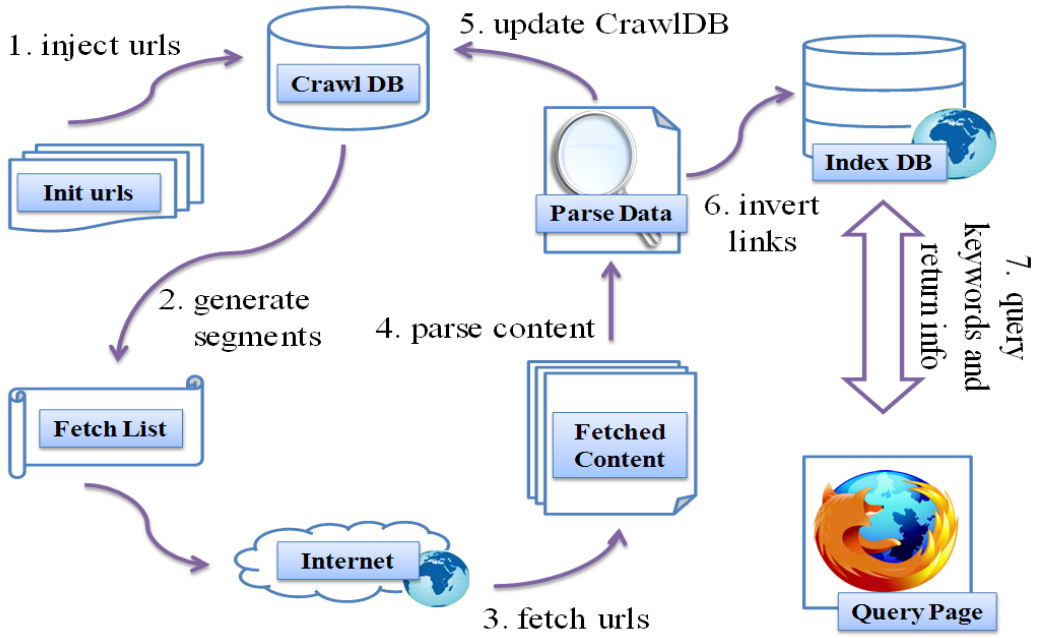


Figure 1. The workflow of Nutch.

Finally, end-users send data to the IndexDB and retrieve the corresponding information via the query page.

By the way, Nutch is very useful but complicated on operating. By using Crawlzilla, user could get much help on operate and manage Nutch.

B. Hadoop

Apache Hadoop [2] is a software framework that supports data-intensive distributed applications under a free license. It enables applications to work with thousands of nodes and petabytes of data. Hadoop was inspired by Google's MapReduce [3] and Google File System (GFS) [4] papers. Hadoop is a top-level Apache project being built and used by a global community of contributors, using the Java programming language. Yahoo! has been the largest contributor to the project, and uses Hadoop extensively across its businesses.

Nutch fetches web pages by MapReduce on Hadoop, which is running on single node by default, but that cannot surfer heavy load. Hence we add cluster calculation into Crawlzilla to balance the load on each compute node when fetching pages are very huge.

C. Search Engine Library

In past experience, the end user only can receive the query result. Search engine just like an mysterious box, input something and search engine will output something. In Nutch, this mysterious box is Lucene, it's an open source project search engine library. When after crawl process and parse website pages, the index pair(url and keywords) will store in search engine library. In private search engine, you can read this mysterious box by some tools, but when you are using Google, you cannot to know the content of Google search engine library. Due to this reason, Crawlzilla provides a tool on the operate website, system administrators can browse on website easily.

III. Related Works and Motivation

Google is a powerful and very userful search engine for many users, but it's not useful in intra website due to it only can search public information. In order to search internal website, system administrators must build search engine by themself or by cost. With Google published search engine system architecture and algorithms, there many research issue in search engine algorithm, file system...etc. Nutch is a complete search engine project, provides crawl algorithm, distributed system, search engine library. For system administrators, to build a private search by Nutch is cumbersome. There are many steps for system install such as install and configure Hadoop, Nutch, Tomcat. After install and configure these services, system administrator must edit crawl website list and download search engine library from HDFS(the file system which is used in Hadoop) and setup the Tomcat, etc., maybe there are some errors during setup Nutch. If the performance isn't very well, the system administrator also need to setup the cluster environment, this is the other cumbersome loop due to build cluster environment is not very easily. For operation, there are many computing services needs to startup/stop, all these operates are in command line, it's not friendly.

In this paper, we developed a toolkit to assist system

 ISBN: 978-1-61208-135-9

administrators to build their search engine not only for install but also for manage and operate. The main core in Crawlzilla is Nutch and the other subsidiary project are frequent update and stable. In addition, we join search engine library API to display the content of search engine library. Due to there different character between English and Chinese words, We also improve the supporting Chinese words, it's can increase the performance of search in Chinese words. We will describe more detail in follow sections. This paper emphasize deploy and operate private search engine and build cluster environment easily, and support more efficiency for Chinese language users. Crawlzilla not focus on improve the performance of Nutch and Hadoop, we focus on operate and friendly use experience.

IV. CRAWLZILLA

In this section, we will introduce the project design concept of Crawlzilla [6] and its architecture. The detail of design concept as the following subsections, we also have several demos and experiments in Section V and Section VI.

A. Overview

Crawlzilla has been released first version called NutchEZ in September 2009. This version used a terminal window interface to help system administrators to submit Nutch jobs, but doesn't provided the other system information clearly. Due to there were many bugs and many function that we can improve in NutchEZ. This release version called Crawlzilla, provides many extend functions, such as support cluster and almost linux base operate system, and more friendly manage interface. The object of Crawlzilla is to provide a search engine tool, witch is easy to install, easy to learn, easy to use and low cost.

1) System Architecture and Design: Crawlzilla divided into three parts, the first part is system installation, the purpose is to integrate and to simplify all of the installation procedures into this phase. The second part is Crawlzilla system management, this tool provides system administrators to check cluster status, set the service of computing nodes(Hadoop cluster process) and web server(Tomcat) and choose language...etc. The third part is the operate interface of search engine management on a website witch is building when system administrators installed Crawlzilla, system administrators can use this website to summit crawl jobs, to manage and browse index pool of search engine, etc. In order to make more efficient search engine operation, it added cluster-type search engine development environment, the main benefits is to increase more efficient in crawling jobs, such as reducing crawl time, support multi task.

2) System Installation: In system installation, we used shell dialog format as the system administrators interface to help system administrators build search engine. By using shell dialog, system administrators don't have to install graphical interface library and provided system administrators to connect remotely via ssh to operate the computing service. In other to provide more friendly interface, installation process will import the system administrators' language automatically, and system administrators can be completed in five steps during the installation process. In addition, in order to allow the installation to the smooth operation of the system after the installation process, the system will create a user account called crawler. Crawler used to start-up all of search engine computing service(e.g., Hadoop, Tomcat...etc), the identity of all the computer cluster will use the same username and password as cluster computing tool of communication between computing nodes. The system design principles to simplify the installation of all the steps, the system administrators just enter a password and select a group to complete the installation of network equipment to meet easy installation design. Cluster installation part, the current system after the installation is designed to produce the relevant Master installation files, the system administrators just copy the installation files in the system to the new computing node can be dynamically installed to add new nodes to the cluster computing.

3) System Service Management: By Nutch and all of search engine components installed, all management must enter the command through the terminal can be implemented, even if the system administrator has successfully installed, if the system administrators are not familiar with the operation of the terminal is still not smooth implementation of the system to search. This paper proposed and developed a system management interface, and currently offers the following functions:

- Check Cluster State: This option provides system administrators to check the current system service state.
- Setup Cluster Computing Service: System administrators can start-up or shutdown the cluster computing service by choose this option.
- Setup Tomcat Service: System administrators can start-up or shutdown the Tomcat server by choose this option.
- Language Switch: Choose English or Chinese version.

In Crawlzilla environment, system need the root permissions to modify the system files(e.g., /etc/hosts) for communication between the cluster computing nodes and system removed during system installation. Crawlzilla will be file to restore the files witch is modified during the installation.

 ISBN: 978-1-61208-135-9

4) System Management: In this section, we only describe the most unique website management system, the major functions provided as follows:

- Crawl:This option provides system administrators to build index pool for search engine. We've simplified most of the commands, the system administrators requires only set their own index pool name, and edit the URL list and the depth of crawling, web page crawling task can be submitted when all of the information will be setting, this feature also supports multiple web crawling, Submit multiple projects simultaneously, improve the utilization of compute nodes.
- Index Pool Management: This option provide system administrators browse the index pool information, such as the initial URL, the local index path, the number of document files and the date of index pool updated ... and other information, the system administrators can delete index pool which is worthless informatio.
- System Status: This option provides system administrators view system status, such as task execution status, the number of clusters ... and so on.

In order to make search engines more flexible, Crawlzilla support mutil search engine. When system administrators have been build index pool, we have to construct the links of existing search engine on the right side of website management system.

V. System Implement

This section will describe how to install and operate Crawlzilla. There are two main functions in Crawlizlla. One is system installation and management; it uses shell script and dialog to implement. The other one is Crawlzilla web management; it use JSP, servlet, Java Bean to implement. Below, it also expresses the detailed process about (1) system installation, (2) system management, (3) crawl setup and index poll management.

A. Installation

User dowlonad Crawlzilla from Google project or Souceforge, then unzips tarball and executes installation file. The install procedure will help system administrators interactively to install and setup. Each step is as follows.

- Step 1: The installation procedure checks system environment and required package.
- Step 2: Creates user "crawler" and setup the password. This user is responsible for Tomcat (web server) service and crawl job submission.
- Step 3: If you see message of Installation completion, go to the URL (http://localhost:8080 or http://your.system.ip.address:8080) to check Crawlzilla web management interface (see Figure 2).

Crawllzilla support single mode and cluster mode. If you just want to install Crawlzilla in one machine, step 1 step 3 is enough. But if you want to build Crawlzilla cluster, you need do step 4 6 in other slaves.

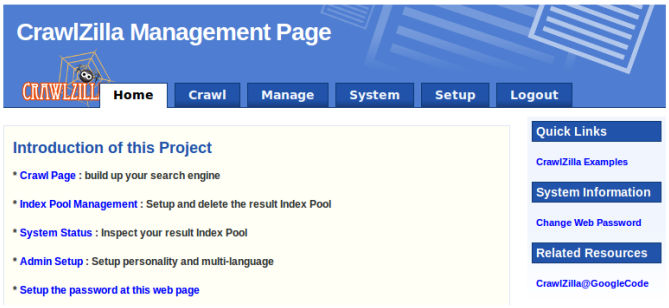


Figure 2. The website of Crawlzilla management, system administrators can use this website to submit the crawling jobs and search engines.

- Step 4: Copies slave installation file from master to slaves.
- Step 5: Executes slave installation file. It will copy some required execution and configuration file from master. It also do authorization with master to make master can assign job to slaves.
- Step 6: Executes system management in master and adds new slave to the Crawlzilla cluster.

B. System management

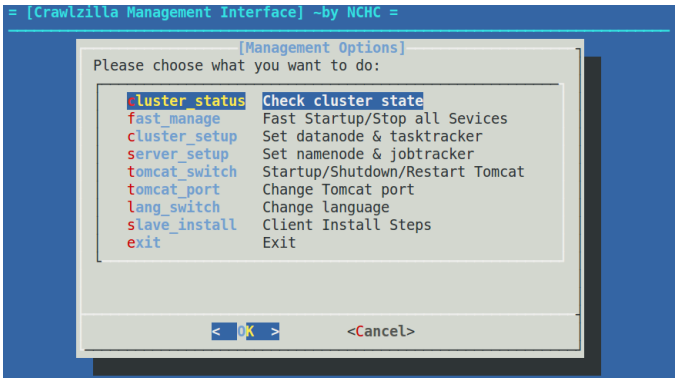


Figure 3. The system management of Crawlzilla, uses can operate the service of computing service by this user interface.

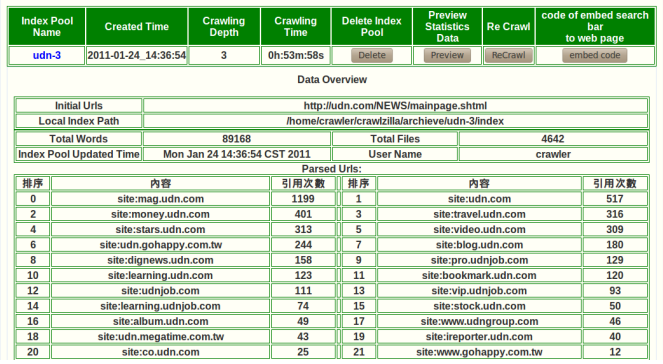
Figure 3 is system management interface, administrator can use this interface to check cluster status, control computing services, control Tomcat web server. It focuses to offers the low-level nodes management compare with Crawlzilla web management. For system administrator, he can use system management to manage cluster and Crawlzilla services through terminal.

C. Crawlzilla web management

System administrators can crawl web data, query status of index pool and manage many search engines through web management. If system administrators wants to crawl some

 ISBN: 978-1-61208-135-9

web data, he just input web URLs, depth and name (the name will be identification for search engine and crawl task).



Index Pool Name	Created Time	Crawling Depth	Crawling Time	Delete Index Pool	Preview Statistics Data	Re Crawl	code of embed search bar to web page
udn-3	2011-01-24_14:36:54	3	0h:53m:58s	Delete	Preview	ReCrawl	embed code

Data Overview

Initial Urls	http://udn.com/NEWS/mainpage.shtml		
Local Index Path	/home/crawler/crawlzilla/archieve/udn-3/index		
Total Words	89168	Total Files	4642
Index Pool Updated Time	Mon Jan 24 14:36:54 CST 2011	User Name	crawler

Parsed Urls:

排序	內容	引用次數	排序	內容	引用次數
0	site:mag.udn.com	1199	1	site:udn.com	517
2	site:money.udn.com	401	3	site:travel.udn.com	316
4	site:stars.udn.com	313	5	site:video.udn.com	309
6	site:udn.gohappy.com.tw	244	7	site:blog.udn.com	180
8	site:dignews.udn.com	158	9	site:pro.udnjob.com	129
10	site:learning.udn.com	123	11	site:bookmark.udn.com	120
12	site:udnjob.com	111	13	site:vip.udnjob.com	93
14	site:learning.udnjob.com	74	15	site:stock.udn.com	50
16	site:album.udn.com	49	17	site:www.udngroup.com	46
18	site:udn.megatime.com.tw	43	19	site:ireporter.udn.com	40
20	site:co.udn.com	25	21	site:www.gohappy.com.tw	12

Figure 4. The information of index-pool, system administrators can read the information of search engine by this page.

When crawling tasks are running, system administrators can click system tab to see operation status in web management. It offers much information (such as process name, disk usage, loading...). When crawling tasks completion, system administrators can see the index-pool name in content of manage tab. It offers analysis function to see how many words, file type, urls in this search engine as shown in Figure 4).

Figure 5. The query screenshot of Crawlzilla search engine.

Figure 5 is query screenshot. Users can use this search engine to query. It will accord index pool to provide query result. The index pool is generated by foregoing crawl task.

VI. Experiment

This section we will propose some experiments to test performance of Crawlzilla. The object of these experiments is to observe the performance with different depth and computing nodes. We will to describe the process and result in the following sections.

A. Experiment Environment

The information of the experiment platform was shown in Table I. We used Core2 Quad CPU with 8GigaBytes RAM to execute test script witch submitted crawling jobs from depth 3 to depth 10. We also to collect the execute time and crawling result whit different computing nodes.

CPU	Intel(R) Core(TM) 2 Quad CPU Q9550 2.83GHz
Memroy	8 GigaBytes
Operation System	Ubuntu 10.04 Lucid(x86)
Crawlzilla Version	0.3.0-101116

Table I
THE PLATFORM OF CRAWLZILLA PERFORMANCE EXPERIMENT.

These experiments will execute in cluster mode (3 computing nodes and 6 computing nodes) and single mode in different depths. It setups the same URL to crawl in Crawlzilla. The result of these experiment as shown in following subsections. The purpose of experiment just for test and verify that the crawlzilla can execute and operate smoothly for general users. It not to show and improve the performance of Nutch and Hadoop.

B. Execute Time

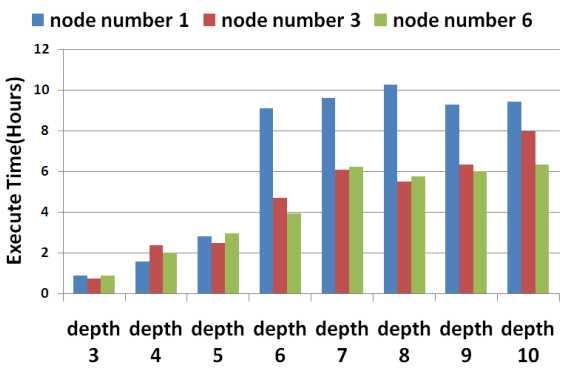


Figure 6. The execute time of Crawlzilla to crawl the same url-list with different depths and computing nodes.

Here are positive indicators of Crawlzilla execute time with different depths and computing nodes, we let Crawlzilla to crawl the same url-list and execute in different environments. Figure 6 shown the result of this experiment. Obviously, we can see the execute time depand on computing nodes. The consumption near depth6 depth 8 are the highest execute time. The execute time also depands on many factors not only computing nodes but ethernet speed and the resources of hardware. Due to Hadoop are useful in process mass data, this result shows if the data not achieve the Hadoop bottleneck, even the system have 10 computing nodes, there aren't enhance the performance significant.

 ISBN: 978-1-61208-135-9

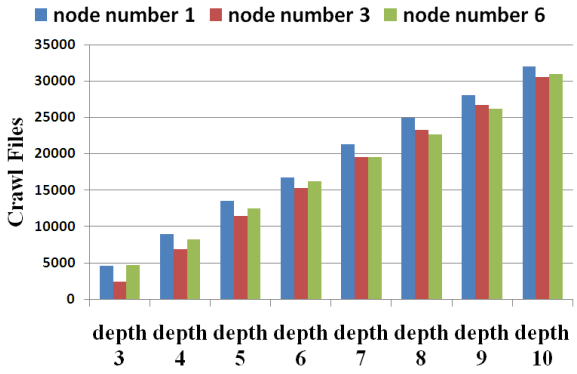


Figure 7. The crawling files with different crawling depths.

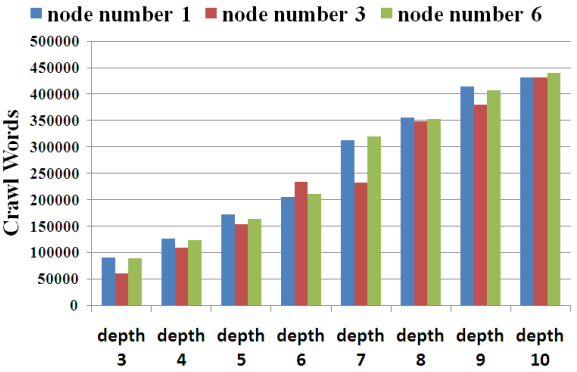


Figure 8. The crawling words with different crawling depths.

C. Crawl Files and Crawl Words

In this part of experiments, we just to test and verify the crawling files and crawling words can grow up with the different depths. As shown in Figure 7, we can see the crawling files in different crawling depths, and the result of crawling words in different crawling depths as shown in Figure 8.

VII. Conclusion and Future Works

In this paper, we proposed and developed a toolkit of search engine, witch is Crawlzilla. It can assist system administrators to build and manage their search engines. Crawlzilla also provides cluster mode, it can improve the performance of crawling jobs. It's a low cost and easy to use toolkit for general system administrators who doesn't have many knowledge in search engine. We also provided several friendly interfaces to assist system administrators to manage and operate their search engines.

The section of experiment, we showed the execute time in different crawling depths whit different computing jobs. According the result, we can observed the cluster mode can improve the crawling performance, obviously. The results of crawling files and crawling words in different crawling depths and computing nodes are nearly, it means the cluster mode can improve the crawling performance and they have the same result.

In future works, we will update this project continually. We will focus on process scheduling of crawling jobs and the other functions. For an interesting phenomenon, we want to use Crawlzilla to observe the six degrees of separation in world wide web. Because the social network like the real world, maybe the six degrees of separation also can completely imitate in world wide web.

References

[1] The Apache Software Foundation, Nutch, available at: http://nutch.apache.org/ , accessed 5 Jan 2011.

[2] The Apache Software Foundation, Hadoop, available at: http://hadoop.apache.org/ , accessed 5 Jan 2011.

[3] J. Dean and S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, In Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6, San Francisco, CA, December 06 - 08, 2004.

[4] S. Ghemawat, H. Gobioff and S. T. Leung, The Google File System, 19th ACM Symposium on Operating Systems Principles, Lake George, NY, October, 2003.

[5] The Apache Software Foundation, Lucene, available at: http://lucene.apache.org/ , accessed 5 Jan 2011.

[6] Crawlzilla Google Code Project Hosting, available at: http://code.google.com/p/crawlzilla/, accessed 15 Jan 2011.

 ISBN: 978-1-61208-135-9