Aggregating product reviews for the Chinese market

YONGLIANG WU

KTH Information and
Communication Technology

Master of Science Thesis
Stockholm, Sweden 2009

TRITA-ICT-EX-2009:208

Aggregating product reviews for the Chinese market

Master thesis

Yongliang Wu
<ylwu@kth.se>

Examiner
Professor Gerald Q. Maguire Jr.

School of Information and Communication Technology
Royal Institute of Technology (KTH)
Stockholm, Sweden

Abstract

As of December 2007, the number of Internet users in China had increased to 210 million people. The annual growth rate reached 53.3 percent in 2008, with the average number of Internet users increasing every day by 200,000 people. Currently, China's Internet population is slightly lower than the 215 million internet users in the United States. [1]

Despite the rapid growth of the Chinese economy in the global Internet market, China's e-commerce is not following the traditional pattern of commerce, but instead has developed based on user demand. This growth has extended into every area of the Internet.

In the west, expert product reviews have been shown to be an important element in a user's purchase decision. The higher the quality of product reviews that customers received, the more products they buy from on-line shops. As the number of products and options increase, Chinese customers need impersonal, impartial, and detailed products reviews. This thesis focuses on on-line product reviews and how they affect Chinese customer's purchase decisions.

E-commerce is a complex system. As a typical model of e-commerce, we examine a Business to Consumer (B2C) on-line retail site and consider a number of factors; including some seemingly subtitle factors that may influence a customer's eventually decision to shop on website. Specifically this thesis project will examine aggregated product reviews from different on-line sources by analyzing some existing western companies. Following this the thesis demonstrates how to aggregate product reviews for an e-business website.

During this thesis project we found that existing data mining techniques made it straight forward to collect reviews. These reviews were stored in a database and web applications can query this database to provide a user with a set of relevant product reviews. One of the important issues, just as with search engines is providing the relevant product reviews and determining what order they should be presented in. In our work we selected the reviews based upon matching the product (although in some cases there are ambiguities concerning if two products are actually identical or not) and ordering the matching reviews by date - with the most recent reviews present first.

Some of the open questions that remain for the future are: (1) improving the matching - to avoid the ambiguity concerning if the reviews are about the same product or not and (2) determining if the availability of product reviews actually affect a Chinese user's decision to purchase a product.

Sammanfattning

I december 2007 uppgick antalet internetanvändare i Kina har ökat till 210 miljoner människor. Den årliga tillväxttakten nådde 53,3 procent 2008, med den genomsnittliga Antalet Internet-användare ökar för varje dag av 200.000 människor. Närvarande Kinas Internet befolkningen är något lägre än de 215 miljoner Internetanvändare i USA Staterna.[1]

Trots den snabba tillväxten i den kinesiska ekonomin i den globala Internetmarknaden, Kinas e-handel inte följer det traditionella mönstret av handel, men i stället har utvecklats baserat på användarnas efterfrågan. Denna tillväxt har utvidgas till alla områden I Internet.

I väst har expert recensioner visat sig vara en viktig del I användarens köpbeslut. Ju högre kvalitet på produkten recensioner som kunderna mottagna fler produkter de köper från on-line butiker. Eftersom antalet produkter och alternativen ökar, kinesiska kunderna behöver opersonlig, opartisk och detaljerade produkter recensioner. Denna avhandling fokuserar på on-line recensioner och hur de påverkar Kinesiska kundens köpbeslut.

E-handel är ett komplext system. Som en typisk modell för e-handel, vi undersöka ett Business to Consumer (B2C) on-line-försäljning plats och överväga ett antal faktorer; inklusive några till synes subtitle faktorer som kan påverka kundens småningom Beslutet att handla på webbplatsen. Uttryckligen detta examensarbete kommer att undersöka aggregerade recensioner från olika online-källor genom att analysera vissa befintliga västra företag. Efter den här avhandlingen visar hur samlade produkt recensioner för en e-affärer webbplats.

Under detta examensarbete fann vi att befintliga data mining tekniker gjort det rakt fram för att samla recensioner. Dessa översyner har lagrats i en databas och webb program kan söka denna databas för att ge en användare med en rad relevanta product recensioner. En av de viktiga frågorna, precis som med sökmotorer är att tillhandahålla relevanta produkt recensioner och bestämma vilken ordning de ska presenteras i. vårt arbete har vi valt recensioner baserat på matchning produkten (men i vissa fall det finns oklarheter i fråga om två produkter verkligen identiska eller inte) och beställa matchande recensioner efter datum - med den senaste recensioner närvarande första.

Några av de öppna frågorna som kvarstår för framtiden är: (1) förbättra matchning - För att undvika oklarheter rörande om Gästrecensionerna om samma produkt eller inte och (2) avgöra om det finns recensioner faktiskt påverka en kinesisk användarens val att köpa en produkt.

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

B2C	Business-to-consumer
B2B	Business-to-business
CFCA	China’s financial security authentication management centre
NPC	the (Chinese) National People’s Congress

1. Introduction

Starting in the early 2000s, economic globalization, trade liberalization, information technology, and business management technology integrated creating modern productive forces. As a result China's E-commerce has been unprecedented in its promotion of the economic vitality of the sector, regional, national, and the world economy -- bring E-commerce to a new level.

China's e-commerce had been through several difficult years of exploration, beginning in 2005. As a result of this exploration its condition improved a lot. By 2007, China's e-commerce infrastructure, planning, regulations, research, and application of theory to practice has made significant progress, due to the rapid development of a modern network infrastructure.

Since 2007, the already remarkable development of computer networks speed up even more. The China-US undersea fiber optic cable enabled an expansion of international internet connectivity. Telephone network growth has also been strong: 160 million fixed telephone subscribers and more than 65 million mobile phones. As a result 13% of the population owns a telephone today. The basic coverage of the national television networks is estimated to provide programming to more than 300 million people [2]. A diverse variety of communications links (optical fiber, microwave, and satellite communications) provides the communications network's backbone. Some of the end products associated with this growth in e-commerce are TV set-top boxes, phones, computers, credit cards, and the Internet.

In addition, regulations for e-commerce have been promulgated; such as those contained in the State Ministry of Information Industry Organization draft entitled “National e-commerce development framework” [3] as part of to be their “Outline for the Development of China's e-commerce strategy” [3]. The Ministry of Information Industry and other units held a symposium for e-commerce laws and regulations in December 1999. In March 2000, China's e-commerce law was proposed in the National People's Congress (NPC) session [2]. Some markets have begun to develop e-commerce in line with local conditions. An important area is security, given the increase in on-line payment - there is a need for antivirus activities to protect users while web surfing. Some parts of China have already introduced regulations, such as the Internet Safety Ordinance [4]. The China Merchants Bank, the Agricultural Bank of China, Bank of China, Industrial and Commercial Bank of China, and others major commercial banks are offering banking services via the internet. The People's Bank of China led the formation of China's financial security authentication management centre [5].

The motivation to do this project on aggregating product reviews is based on a simple story: In 2008, I wanted to buy a Sony Ericsson mobile phone in Stockholm supports Chinese SMS, not all phones have this function in Sweden. Before I bought anything in store, I used Google (as everyone else does) to search for the product specially the website provides expert/user reviews on it. There were many English reviews out there among very few Chinese reviews, which I wanted know my Chinese fellows' personal experiences after they bought the phone I wanted. I was a bit frustrated that it took too much time to browser few hundred pages on Google to find what I really needed. What I needed was just a simple website listed all the related reviews collected from different users or sources to tell me what is good and bad for a product. This is where the motivation comes from. I smelt fragrance from China's online product review market cake.

Following the introduction in this chapter, Chapter 2 provides background about product reviews, data mining, importance of product reviews, and aggregation of product reviews. Chapter 3 introduces the method we will use for aggregating reviews, while Chapter 4 describes the organization of a database to store the reviews. Chapter 5 describes how we used data mining to collect product reviews and insert them into the database. Given a collection of reviews produced by the data mining, Chapter 6 describes how this aggregated data can be presented to users and how we evaluated the success of our method. Finally, Chapter 7 summarizes our findings and suggests some future work. The appendices contain the source code for inserting products into the database and the code for inserting reviews into the database.

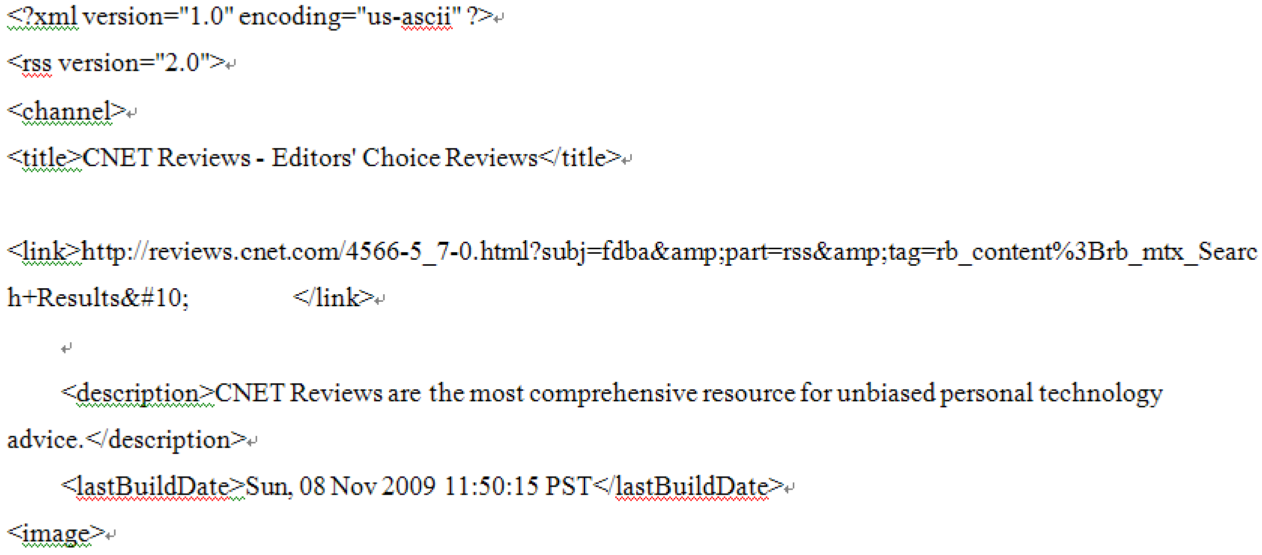
2. Background

2.1 Product reviews

There are two types of product reviews on the internet. One type is written by a website owner who writes their own reviews of a product for their readers. Most of time, the website owner is doing this to increase the return on their investment. For instance, if you click a link on this website and buy something, the owner will get a percentage of the sale or will get paid by how much traffic this website sends to a specific shopping site.

The second type of product review is user generated reviews. In this type of review, customers review products and publish them on a website. People tend to trust this type of review more, as the reader believes that someone else really used the products and is presenting the advantages and disadvantages from the point of view of an actual user. This type of review may provide a potential customer for this product with the information that they need to make their purchase decision. Note that these reviews are typically not written as part of for profit service, although in some cases the author of the review may receive compensation for their reviews (see for example Amazon.com's Associates program - http://affiliate-program.amzaon.com).

Figure 1 is an example of a standard product review format from cnet.com. This format could be used by many websites.



```
<?xml version="1.0" encoding="us-ascii" ?>
<rss version="2.0">
<channel>
<title>CNET Reviews - Editors' Choice Reviews</title>

<link>http://reviews.cnet.com/4566-5_7-0.html?subj=fdba&amp;part=rss&amp;tag=rb_content%3Brb_mtx_Searc
h+Results&#10;           </link>

    <description>CNET Reviews are the most comprehensive resource for unbiased personal technology
advice.</description>
    <lastBuildDate>Sun, 08 Nov 2009 11:50:15 PST</lastBuildDate>
<image>
```

Figure 1: An example product review format from cnet.com [6]

2.2 Data mining

In order to collect a large number of products reviews one either needs a lot of employees to write product review (the first type of review discussed in the previous section) or you need to have lots of non-employees writing such reviews. In the second case, these non-employees can either be explicitly writing these reviews for your site or they could be writing them for others site or even just reviewing the product in their blog or on their personal web site. This thesis will focus on the second type of reviews - particularly the case where the reviews have not been written expressly for this site. This implies that we need to (1) find these reviews and (2) collect them. This process is known as data mining.

Data mining is the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform these data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery. [7]

The legal status of data mining of other sites in China and the legal status of using someone else's reviews is “为介绍、评论某一作品或者说明某一问题，在向公众提供的作品中适当引用已经发表的作品；可以不经著作权人许可，不向其支付报酬” [8]. Translated into English, this means that a comment, introduction, or explanation of a work to the public with an appropriate quote; can be done without requiring the copyright owner’s permission, if no payment involved. Thus the original review sources can be listed on a website, provided that their source is appropriately noted and that no payment is involved. Thus it would appear that a non-profit use can be made of the reviews of others. However, a for-profit use would require the copyright owner's permission.

2.3 Importance of product reviews

Customer reviews have an important influence on the purchases of the on-line shoppers; particularly for customers who need to know more about the product. Earlier studies of on-line shops that have examined the role of the customer comments have indicated that users, comments at the on-line retail site cannot be ignored, particularly in the areas of certain products. These reviews may even be a significant basis for a purchase decision. Examples of studies indicating the importance of these reviews are given in the next paragraph.

The rational use of expert/user reviews by on-line shops can promote their on-line sales. Research carried out by eVOC Insights, LLC in 2006 on “Ratings, Reviews and the Customer Decision Process: Amazon v. BestBuy v. CircuitCity v. Walmart” shows “63% of consumers indicate they are more likely to purchase from a site if it has ratings and reviews” [9]. One of the key results of a University of Omaha study is that “customer reviews had a statistically significant effect on increasing sales at NetShops, driving a 36% increase in sales of product with reviews over this period, with a 95% confidence level” [10]. Other research indicates that 9 percent of on-line shoppers will be writing of their favorite on-line shopping experience by writing product reviews, but 4 percent of the shoppers say they do not like to write product reviews. If the consumer’s shopping experience is enjoyable, then 43 percent of customers will be relatively happy to leave their comments for other shoppers, but an unpleasant shopping experience results in only 17 percent of the people being willing to write a review. Shoppers commented that users found the usefulness of the product reviews vary, but most value user reviews of consumer electronics and computers, followed by books, software, music, and DVDs. However, expert/user reviews and ratings is a trend, with expert/user reviews and ratings second in importance to search results [11].

An expert review in the expert’s area has greater influence on a customer’s purchase decision than an ordinary customer’s review of the same product. The role of expert/user reviews cannot be ignored, particularly for certain products, and may even be the most important basis of a purchase decision [12]. In this regard, on-line shoppers will repeatedly make their purchases on sites that provide a good rating and review web site. Customer’s behaviors demonstrate that e-commerce website product reviews have a value in the customer's purchase decision-making. The impact of various factors on their purchase decision ranks the most important factors (in descending order) as: competitive prices, detailed product descriptions, user-friendly web site, good customer service, rich variety of choice, feature comparison and a clear picture of the goods, the brand name rating, expert/customer reviews, and buying guide.

2.4 Importance of aggregating product reviews

It is important how many reviews there are, as according to Powerreview.com: “74 percent want to read a minimum of between 2-7 customer reviews per product to have sufficient confidence to judge a product. 63 percent want to read more; specifically between 4-15 reviews” [13].

In the following chapter we will present our method for producing a website with a large number of product reviews. As more reviews would attract more visitors, we need to understand how potential visitors will find the site. Attracting visitors to the site depends on search engines, especially Google. Thus it is important to produce a web site that is (1) visible to the appropriate search engines and (2) to produce information (content) that will be highly ranked. Of course there is a strong feedback effect as if there are highly ranked pages, this high ranked content will bring more visitors, more visitors potentially means more buyers, ultimately leading to more user written reviews - which leads to more fresh reviews and higher page ranking.

Today, there is at least one Chinese product review aggregator called ksou.com.cn. They organize all their reviews using automatic classification in order to provide the customer with comprehensive product reviews. They claim on their website that all information on their site is automatically generated, i.e., there is no manual editing or processing. This claim leads to one of our criteria for success, minimizing the amount of manually produced material -- as needing to manually produce material decreases the number of items that we can provide reviews for - hence potentially reducing our page rank. Minimizing the amount of manual work that is required is also important to minimize costs.

3. Method

The primary practical goal of this thesis project is to build a web site offering Chinese product reviews from different Chinese sources, plus some English language product reviews for the demonstration purposes. Ideally we want to aggregate as many review sources in order to provide the best possible collection of products reviews for potential readers. In this project we will aggregate 20 Chinese review sources and 10 English language review sources. These sources are shown in Table 1 and Table 2.

Table 1: Chinese review sources

http://notebook.ccw.com.cn
http://www.3qit.com
http://notebook.pconline.com.cn
http://www.it.com.cn
http://eva.139shop.com
http://www.pcpop.com
http://www.21tx.com
http://digi.tech.qq.com
http://www.pjtime.com
http://digi.tech.com
http://wangyou.pcgames.com.cn
http://www.gamespot.com.cn
http://gameonline.yesky.com
http://www.enet.com.cn
http://sc.cbinews.com
http://dvdc.thethirdmedia.com
http://www.hi-pda.com
http://mobile.intozgc.com
http://www.pj.com.cn
http://homea.people.com.cn

Table 2: English review sources

http://www.ciao.co.uk
http://www.cnet.com
http://www.pcworld.com
http://www.pcmag.com
http://www.dpreview.com
http://www.engadget.com
http://www.dcresource.com
http://www.gamespot.com
http://www.pcphotomag.com
http://www.phonedog.com

The sources listed in table 1 and 2 are very popular review websites for people to either follow the latest trends or gather information before make purchase decisions. Some of the English sources have a very Alexa rank, for instance cnet.com, pcworld.com, and pcpop.com.

3.2 Technique and tools to be used

A tool called screen-scraper will be used to aggregate product reviews from different Chinese product review sources - as we have not been able to find any Chinese websites using Web 2.0 technology for their reviews (details of our data mining approach are given in chapter 5). MySQL 5.0 will be used as a data base to store the product reviews (details of the database will be given in the next chapter). Note that a screen-scraper is a tool for data mining information from web sites, even if the sites have different structure.

WampServer, a Windows web development environment, will be used for development as “It allows you to create web applications with Apache, PHP, and the MySQL database. It also comes with PHPMyAdmin and SQLiteManager to easily manage your databases.” (See more info: http://www.wampserver.com). The combination of LAMP (linux, Apache, PHP/Perl, and MySQL) could also have been used (for details see O'Reilly Media, Inc.'s ONlamp site: http://onlamp.com/).

It is important that we respect the rights of the web sites that we might potentially data mining for content. One aspect of this is to avoid mining sites that have explicitly said that they do not want web spiders or robots (programs that automatically crawl web space) to access their site (this is according to The Robots Exclusion Policy). The site does this by placing a file ”robots.txt” at the site. Well behaved robots should respect the directives found in this file. For details see [14],[15], and [16]. We will obey the no-robots directive of both web pages and directories (For details of the "no-robots" mechanisms see chapter 6).

4. Database design

This section descries the database structure. There are three tables: (1) products table – shown in Figure 3, (2) reviews table – shown in Figure 4, and (3) sources table – shown in Figure 5. The relationship of all three tables is shown in Figure 2.

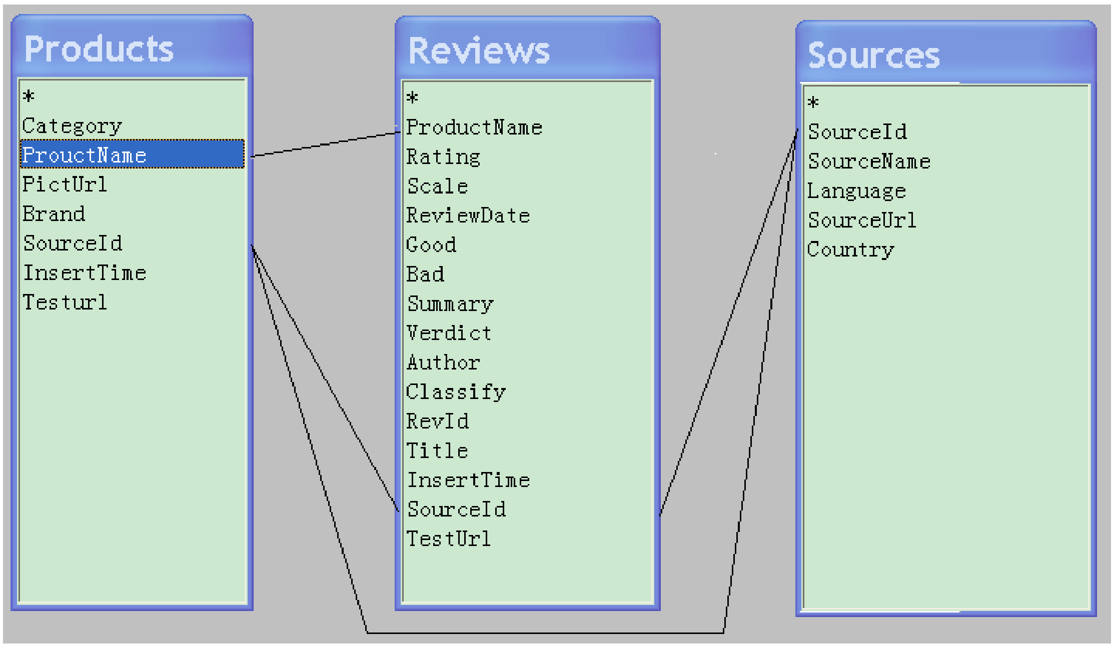


Figure 2: Database structure

This database structure was chosen to make it easy to generate the content for a review web site. As the focus is reviews of products, it was natural that information about a specific product is stored together in the "Products" table. For each product in this table, there can be zero or more product reviews. Thus the reviews are kept in their own table. In order to provide further information about a given product review we also remember where each review came from - to that we can generate a link to the original content. Details of each of these tables will be described in the following sections.

4.1 Product table

The products table shown in Figure 3 contains basic information about a product: Category, ProductName, PicUrl, Brand, SourceId, Inserttime, and Testurl. The primary keys are ProductName (a variable length string) and SourceId (an integer). ProductName is linked to the Reviews' table ProductName. SourceId is linked to the Reviews' table SourceId and the Sources' table SourceId. Note that the sourceId is simply an opaque integer and the value has no external meaning; this is purely an internal identifier that is used to tie together products, reviews, and sources.

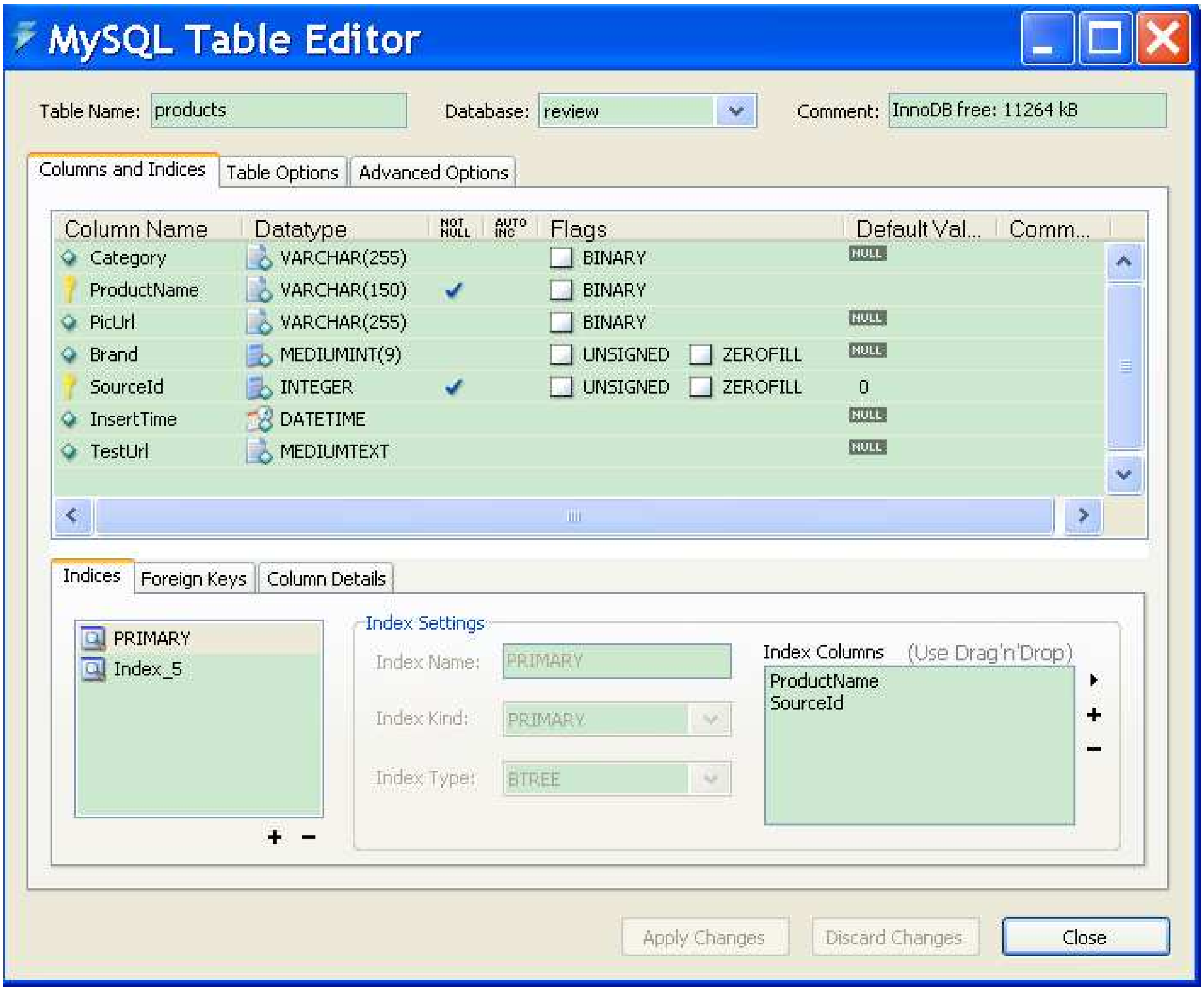


Table Name: products Database: review Comment: InnoDB free: 11264 kB

Column Name	Datatype	NOT NULL	AUTO INC	Flags	Default Val...	Comm...
Category	VARCHAR(255)			□ BINARY	NULL	
ProductName	VARCHAR(150)	✓		□ BINARY		
PicUrl	VARCHAR(255)			□ BINARY	NULL	
Brand	MEDIUMINT(9)			□ UNSIGNED □ ZEROFILL	NULL	
SourceId	INTEGER	✓		□ UNSIGNED □ ZEROFILL	0	
InsertTime	DATETIME				NULL	
TestUrl	MEDIUMTEXT				NULL	

PRIMARY
Index_5

Index Settings
Index Name: PRIMARY
Index Kind: PRIMARY
Index Type: BTREE

Index Columns
ProductName
SourceId

Figure 3: Products table

The products table creation syntax is:
CREATE TABLE `review`.`products` (
`Category` varchar(255) COLLATE utf8_unicode_ci DEFAULT NULL,
`ProductName` varchar(150) COLLATE utf8_unicode_ci NOT NULL,
`PicUrl` varchar(255) CHARACTER SET utf8 DEFAULT NULL,
`Brand` varchar(45) COLLATE utf8_unicode_ci DEFAULT NULL,
`SourceId` int(10) NOT NULL,

```
`InsertTime` timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP,
`TestUrl` mediumtext CHARACTER SET utf8,
PRIMARY KEY (`ProductName`,`SourceId`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_unicode_ci;
```

4.2 Reviews table

The reviews table shown in Figure 4 has ProductName (a variable length string), Rating (a variable length string), Scale (a variable length string), ReviewDate (a variable length string), Good (a medium length string), Bad (a medium length string), Summary (a medium length string), Verdict (a medium length string), Author (a variable length string), Classify (a variable length string), RevId (an integer), Title (a variable length string), InsertTime (current_timestamp), SourceId (an integer), and Testurl (a medium length string). The primary key is RevId (an integer). The ProductName is linked to the Products' table ProductName, the SourceId is linked to the Products' table SourceId and the Sources' table sourceId.

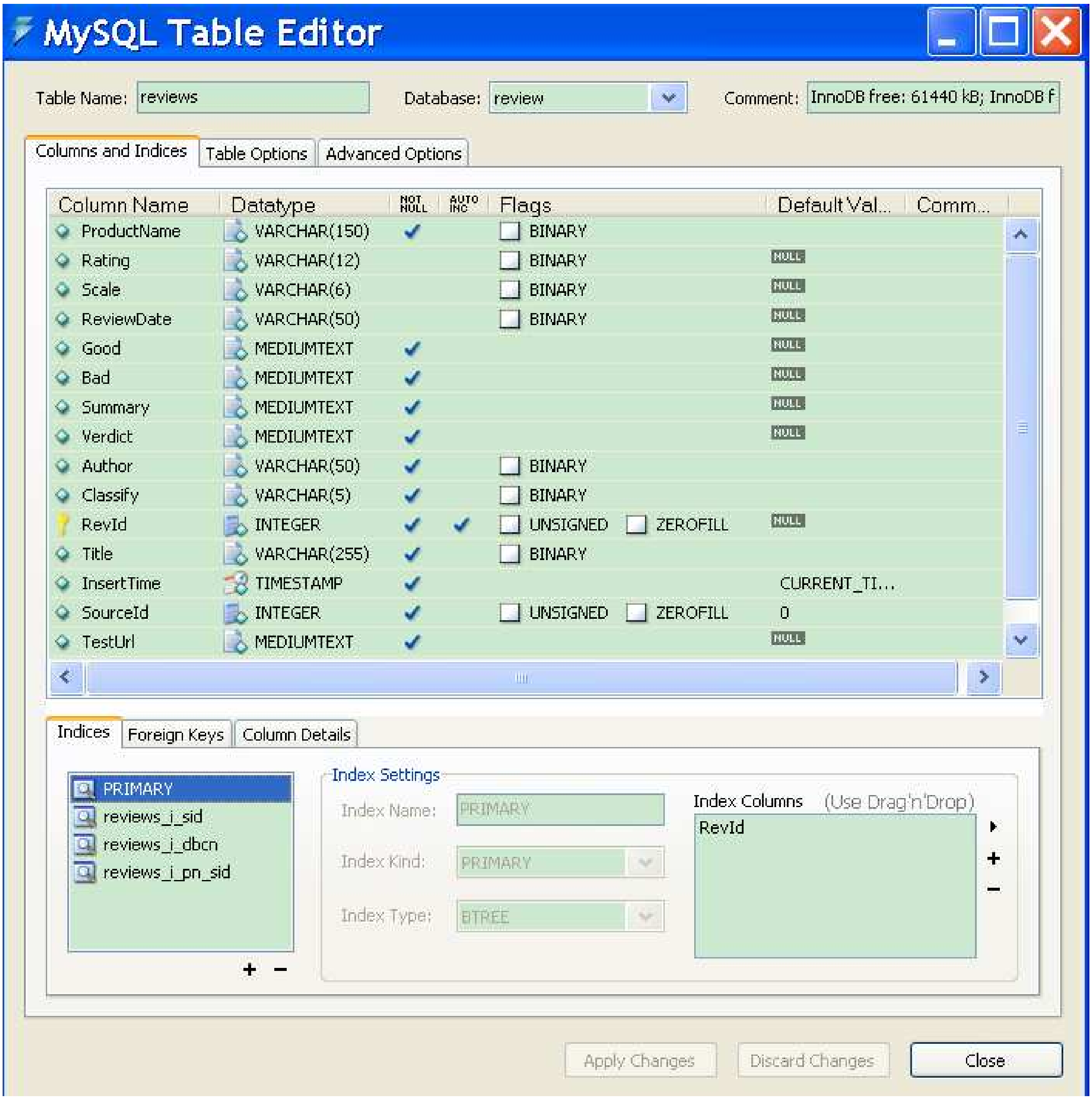


Table Name: reviews   Database: review   Comment: InnoDB free: 61440 kB; InnoDB f

Column Name	Datatype	NOT NULL	AUTO INC	Flags	Default Val...	Comm...
ProductName	VARCHAR(150)	✓		☐ BINARY		
Rating	VARCHAR(12)			☐ BINARY	NULL	
Scale	VARCHAR(6)			☐ BINARY	NULL	
ReviewDate	VARCHAR(50)			☐ BINARY	NULL	
Good	MEDIUMTEXT	✓			NULL	
Bad	MEDIUMTEXT	✓			NULL	
Summary	MEDIUMTEXT	✓			NULL	
Verdict	MEDIUMTEXT	✓			NULL	
Author	VARCHAR(50)	✓		☐ BINARY		
Classify	VARCHAR(5)	✓		☐ BINARY		
RevId	INTEGER	✓	✓	☐ UNSIGNED ☐ ZEROFILL	NULL	
Title	VARCHAR(255)	✓		☐ BINARY		
InsertTime	TIMESTAMP	✓			CURRENT_TI...	
SourceId	INTEGER	✓		☐ UNSIGNED ☐ ZEROFILL	0	
TestUrl	MEDIUMTEXT	✓			NULL	

- PRIMARY
- reviews_i_sid
- reviews_i_dbcn
- reviews_i_pn_sid

Index Name: PRIMARY
Index Kind: PRIMARY
Index Type: BTREE
Index Columns (Use Drag'n'Drop): RevId

Figure 4: Reviews table

The reviews table creation syntax is:

```
CREATE TABLE `review`.`reviews` (
`ProductName` varchar(150) COLLATE utf8_unicode_ci NOT NULL,
`Rating` varchar(12) COLLATE utf8_unicode_ci DEFAULT NULL,
`Scale` varchar(6) COLLATE utf8_unicode_ci DEFAULT NULL,
`ReviewDate` varchar(50) COLLATE utf8_unicode_ci DEFAULT NULL,
`Good` mediumtext CHARACTER SET latin1,
`Bad` mediumtext CHARACTER SET latin1,
`Summary` mediumtext COLLATE utf8_unicode_ci,
`Verdict` mediumtext COLLATE utf8_unicode_ci,
`Author` varchar(150) COLLATE utf8_unicode_ci DEFAULT NULL,
`Classify` varchar(5) COLLATE utf8_unicode_ci DEFAULT NULL,
`RevId` int(11) NOT NULL AUTO_INCREMENT,
`Title` varchar(255) COLLATE utf8_unicode_ci DEFAULT NULL,
`InsertTime` timestamp NOT NULL DEFAULT CURRENT_TIMESTAMP,
`SourceId` int(10) NOT NULL,
`TestUrl` mediumtext COLLATE utf8_unicode_ci,
PRIMARY KEY (`RevId`)
)ENGINE=InnoDB AUTO_INCREMENT=3232 DEFAULT CHARSET=utf8
COLLATE=utf8_unicode_ci;
```

4.3 Sources table

The Sources table shown in Figure 5 includes SourceId (an integer), SourceName (a variable length string), Language (a variable length string), SourceUrl (a variable length string), and Country (a variable length string).

This table describes the information for a source. The primary key is SourceId and the SourceId is linked to the Products' and the Reviews' table SourceId.

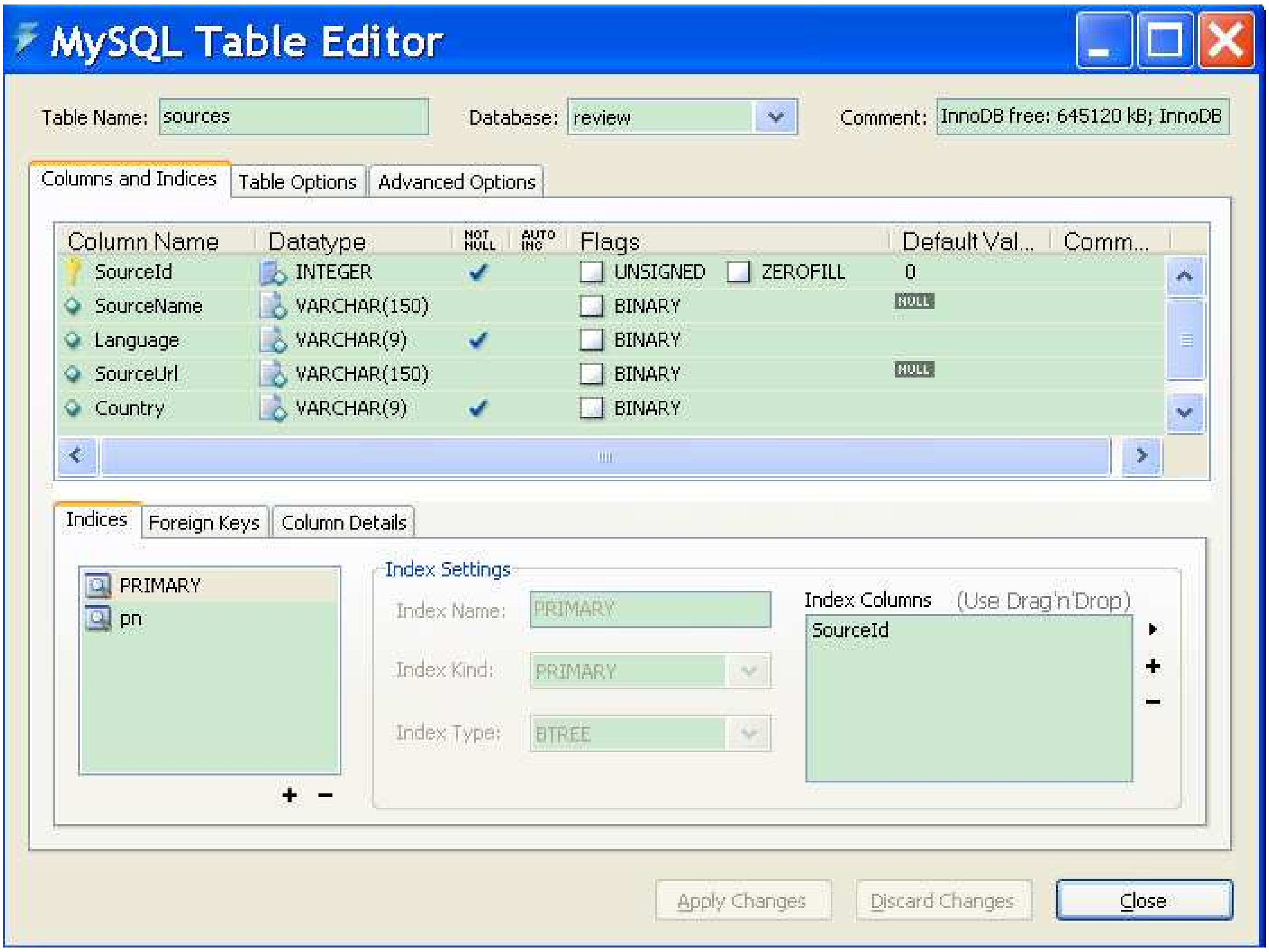


Table Name: sources

Database: review

Comment: InnoDB free: 645120 kB; InnoDB

Column Name	Datatype	NOT NULL	AUTO INC	Flags	Default Val...	Comm...
SourceId	INTEGER	✓		□ UNSIGNED □ ZEROFILL	0	
SourceName	VARCHAR(150)			□ BINARY	NULL	
Language	VARCHAR(9)	✓		□ BINARY		
SourceUrl	VARCHAR(150)			□ BINARY	NULL	
Country	VARCHAR(9)	✓		□ BINARY		

PRIMARY
pn

Index Settings

Index Name: PRIMARY

Index Kind: PRIMARY

Index Type: BTREE

Index Columns (Use Drag'n'Drop)

SourceId

Figure 5: Sources table

The sources table creation syntax is:

```
CREATE TABLE `review`.`sources` (
`SourceId` int(10) unsigned NOT NULL,
`SourceName` varchar(150) CHARACTER SET utf8 COLLATE utf8_unicode_ci
NOT NULL,
`Language` varchar(9) CHARACTER SET utf8 COLLATE utf8_unicode_ci
DEFAULT NULL,
`SourceUrl` varchar(150) CHARACTER SET utf8 COLLATE utf8_unicode_ci
DEFAULT NULL,
`Country` varchar(9) CHARACTER SET utf8 COLLATE utf8_unicode_ci DEFAULT
NULL,
PRIMARY KEY (`SourceId`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_bin;
```

5. Data mining design

You may never have heard of data mining, but if you have Internet access, you are likely to be a beneficiary of information retrieval methods. Today many programs are used to retrieve information in an orderly manner with specialized tools. One approach to mining web based information is to imitate a web browser with a program that sends HTTP requests and parses the response. This software can make use of knowledge about the layout of different sorts of web pages to collect value data. Information from such data mining can be used for a wide variety of purposes. For example, the information might be stored into a database to collect historic information (for example, the arrival times of planes at an airport, the temperature as reported by a specific weather station ...). Alternatively the information might be used immediately, to provide user with a notification of the impending arrival of a plane that they are waiting for or the current weather. Additionally current and historic information can be combined to automatically execute a stock trade given the current price and historic price and volume information. Similar application exists in many domains, such as real estate, finance, meteorology, public transportation, etc.

5.1 Screen scraping

To collect reviews from different websites into my central database, we use a data mining technique called screen scraping (in this particular content the process is also referred to as web scraping) [17]. The mining tool that is used is call screen-scraper. This application is a user friendly tool.

Here are some of the advantages and disadvantages by using screen-scraper for data extraction:
Advantages:

1) This application avoids needing to learn the details of regular expressions, the HTTP protocol, about application cookies, etc. Thus lowering the effort needed to start collecting reviews.

2) Once you learn a particular screen-scraping applications, you can rapidly adapt it to crawl other web sites

3) As it is a commercial product, the company offers technical support

Disadvantages:

1) Unfortunately, each screen-scraping application has its own way of

doing things. This leads to a high learning curve and significant lock-in to a given screen scraping tool. This means that you have to commit rather early to select a specific screen scraping tool and you can not easily change to another tool.

2) A potential cost. This is commercial software required to pay fees to use the full version. (There are three versions: Basic, Professional, and Enterprise - with prices of Free, US$499, US$2,499 - according to http://www.screen-scraper.com/download/choose_version.php.)

There are also other ways for data extraction. Software called "iMacros" (http://www.iopus.com/imacros/) can do the similar thing if you already familiar with regular expressions and at least one programming language. Sometimes, it can be very complex and painful for those who do not have a lot of experiences with regular expressions.

Screen-Scraper comes in three different editions: Basic, Professional, and Enterprise. We have used a trial version of screen-scraper Enterprise Edition. There are some differences between the three. For example the "Send email from a script" is only functioned in Enterprise Edition not the other two editions [18]. The trial version is fully-functional for 30 days. As noted previously we try to obey the relevant regulatory rules and the individual website robot exclusion policies (see section 6.1). The biggest advantage of the screen-scraper application is its ease of use. This is also a widely used solution. However, if you do not mind paying a bit more money, you can save a lot of time in return. If you want to scrape a single web page, you can get what you want by writing regular expressions in almost any programming languages you like in a short time. However, if you want to crawl hundreds of thousands of different structures of web sites, it is advantageous to use one of the specially designed web scraping applications.

I followed the tutorials on screen-scraper community and they have also a tutorial in video format which did help me step-by-step through the process generally used to scrape information from web pages using screen-scraper (http://community.screen-scraper.com/Tutorial_1_Page_1).

5.2 Coupling the screen scraping tool to the database

I have created two interpreted Java scripts to generate data and store it into a database. These scripts are called "product to db" and "review to db", these files insert scraped information into the tables in MYSQL database. The complete scripts can be found in Appendix A and B.

6. Analysis

6.1 Aggregation ethics

The Internet can be considered to be the world's largest database, because it provides access to databases, files, and computer systems. Today the major search engines (such as those of Google, Microsoft, Yahoo!, etc.) enable users to simply type in some words and retrieve a (potentially) long list of items that are related to these words. Because the Internet is a valuable resource, many entrepreneurs are seeking new and innovative ways to build value based upon this information. This sometimes causes conflicts between those who have information and those who want to use this information in new ways. These conflicts concern both ethical and legal concerns. We will see in the following paragraphs examples of both types of concerns.

The owner of a site benefit from advertising the contents of his/her Web site. Usually site owners try (and pay) to publicize their products and/or services. So why would a site owner not want his or her site scraped? Why do site owners may not want others to crawl their web site's data? Why would some site owner's block the web robots of some or all search engines? Some people think that accessing web sites (other than through a browser) is data mining. Why would a web site restrict access to just browsers or even browser's from specific sites? Part of the answer lies in the site owner wanting to be able to count or identify the number of users accessing this information - for example, this might be important for their marketing and operations budgeting. Others might want to ensure that the user sees their content in a controlled manner, for example only when presented with their own layout - perhaps with their own advertisements. Others might want to restrict access to this information to those who have an IP address within some specific range (for example, to provide different information to different users).

It should be noted that there are some web sites that contain public information that can be re-purposed almost without limits - generally requiring that the source be acknowledged. While other sites place greater limits on re-purposing of their content. While some site prohibit any use other than via browsing their web site.

One of the original reasons for sites to prohibit or restrict the web robots from crawling their site is that such robots place an additional burden on the server. To limit the bandwidth and server capacity used, a polite web robot will limit when and how much it accesses a web site; thus it might crawl only part of the web site at a given time and repeatedly access this site at different times in order to collect the desired information. In addition to limiting when and how much access is made to a web site, polite web robots can also adjust their web crawling to minimize their impact upon a web site by changing the rate at which they make queries based upon the inverse of the service time (i.e., if a site is heavily loaded, thus the delay for responses is high - then the crawler slows down or at some point avoids crawling this site). In any case all robots should obey the specifications of the robots.txt file of a site.

Two additional aspects of using content from a web site is that you need to ensure that your use does not infringe upon the copyright owner's rights and that you are in compliance with the site's terms of service. Unfortunately, ensuring compliance with copyrights and terms of service requires human intervention - thus a human must decide whether to add or not add a given site to the list of sites that you are going to crawl to generate aggregated product reviews. Note that there are some initiatives to use machine readable policy documents to enable this process to be automated, but the details of this are outside the scope of this project. In this thesis project we have manually selected the sites listed in tables 1 and 2 (on page 7) and read the sites' terms of service to ensure that are use is in compliance with these terms.

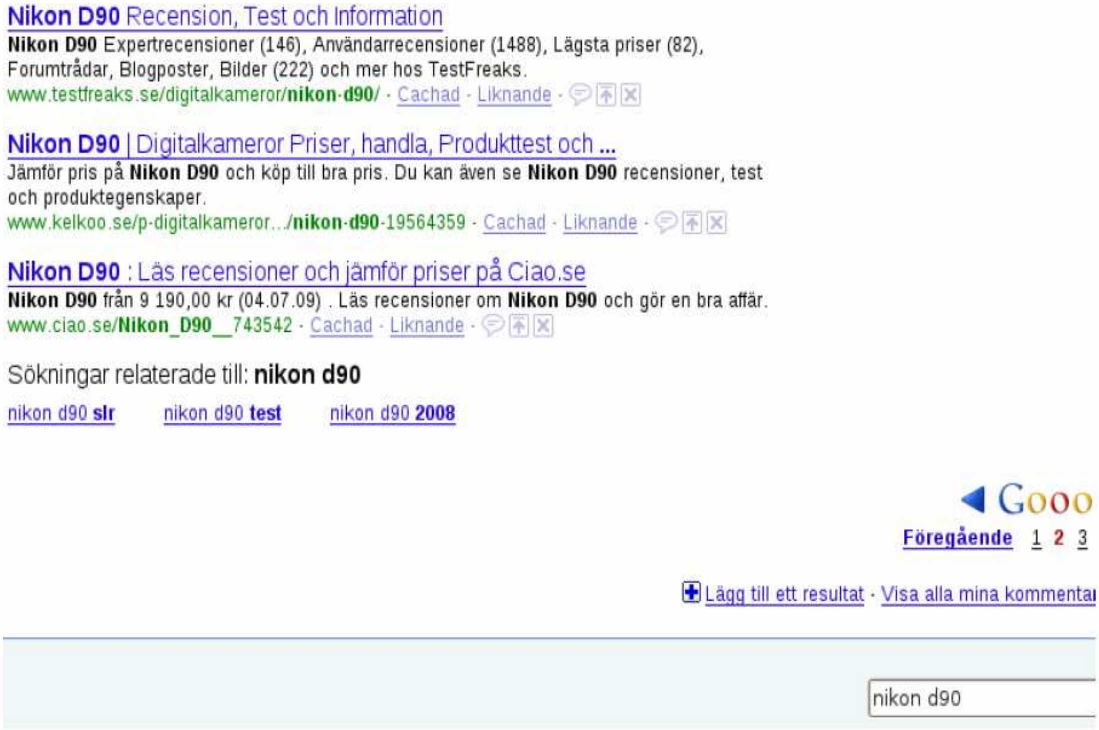
6.2 How much of the site content needs to be manually generated

As manually generated content increases the cost of operating a service, it is desirable to minimize this content. However, as noted above there is some human interaction required to identify which sites the automated tools will craw for content. Once these sites are identified there is relatively little that the user must do -- besides error control. Error control is necessary as (1) sometimes sites may change their layout or policies and (2) sometimes the web scraping program is not able to process one or more items. However, we should keep in mind that we want to provide a large number of reviews to help on-line shoppers to make their decisions; thus we must do the amount of work necessary to achieve a critical mass of reviews. In this project we have used a limited number of web sites (see chapter 3) - but a more extensive set would be needed for a service to be made available to the public.

6.3 How high a ranking can an aggregated content website achieve

A higher ranking on search engines, especially on Google, can attract a lot of traffic to a site. As can be seen in Figure 6, an aggregation website (in this case testfreaks.se) has a higher ranking than ciao.se and kelkoo.se when we search for the product “Nikon d90”. The traffic rank of testfreak.com is 7818 on alexa.com on 08 Nov. 2009.The ranking is in terms of rank order starting with the highest traffic site

(Google.com has rank 1) and the Chinese player ksou.com.cn ranks only 235945 at alexa.com on Nov Dec. 2009 [19]; thus there is a long way to go for Chinese aggregators.



Nikon D90 Recension, Test och Information
Nikon D90 Expertrecensioner (146), Användarrecensioner (1488), Lägsta priser (82), Forumtrådar, Blogposter, Bilder (222) och mer hos TestFreaks.
www.testfreaks.se/digitalkameror/**nikon-d90**/ - Cachad - Liknande

Nikon D90 | Digitalkameror Priser, handla, Produkttest och ...
Jämför pris på **Nikon D90** och köp till bra pris. Du kan även se **Nikon D90** recensioner, test och produktegenskaper.
www.kelkoo.se/p-digitalkameror.../**nikon-d90**-19564359 - Cachad - Liknande

Nikon D90 : Läs recensioner och jämför priser på Ciao.se
Nikon D90 från 9 190,00 kr (04.07.09) . Läs recensioner om **Nikon D90** och gör en bra affär.
www.ciao.se/**Nikon_D90**__743542 - Cachad - Liknande

Sökningar relaterade till: **nikon d90**

nikon d90 **slr** nikon d90 **test** nikon d90 **2008**

Figure 6: Search result of Nikon d90

.

6.4 What barriers are there to creating a successful aggregated review website

Key barriers to creating a successful aggregated review web site include: (1) providing timely aggregation, i.e., minimizing the time between a new review becoming available and this review being included in your aggregate reviews; (2) achieving a critical mass of reviews - so that it is worthwhile for visitors to come to your site rather than search for reviews themselves - for example, using a search engine; (3) the aggregated site has to have high enough performance that user's have a good experience when visiting the site (i.e., the web site has to scale well with increasing number of users) [20]; (4) the site needs high availability - as if users are not able to access the site they will soon stop trying to access it; (5) users need to have confidence in the accuracy of the results that they find - for example, the aggregated reviews should be balanced and not simply limited to favorable reviews. An important aspect of success is the general market for e-commerce, since if the users can not afford to buy products on-line or they experience poor products (or support) they will not be interested in making new purchases on-line. While there is currently a global recession there are some signs for continued growth in e-commerce. As of July 6, 2009, AliPay (www.alipay.com) reached 200 million registered users, on volume of 700 million transactions. Although the Internet has become a mainstream consumer consumption patterns as in the fourth quarter of 2008 reveal that China's network sales accounted for only 1.39 percent of total retail sales of social commodities, while over the same period the United States on-line sales reached a level of 6% and South Korea 14%. One way of interpreting these numbers is that the Chinese e-market potential is huge. [21]

Some analysts predict that the number of Chinese Internet users will increase from 30 million to 60 million in two years (China Internet Network Information Center), suggesting that the amount of Internet consumption will increase at that time – resulting in the Internet becoming truly integrated into people's lives. As economic conditions improve in China (and elsewhere), there is a chance for increased e-commerce spending as the potential purchasers will have money left over after paying their daily living expenses, housing, etc. If so, it may be feasible for a review aggregation site to attract eyeballs by purchasing Google Ad sense words to get visitors. [22]

6.5 How does my solution compare to the solution of others, specifically ksou.com.cn or Amazon.com

The business model for ksou.com.cn is pure Business-to-consumer. Unlike amazon.com, we are not trying to sell anything to my visitors, but rather we are trying to facilitate their purchase decision. Thus is my core business is aggregating reviews and selling access to these aggregated reviews to online-shops (who will use this information to help their visitors to make purchase decisions -- hopefully via their own sites). My aggregated reviews could be combined with information from price comparison sites, enabling a win-win situation. Thus the business model that we are adopting is similar to that of Consumer Reports (http://www.consumerreports.org) in providing aggregated reviews and **not** being a party to the commercial purchase or sale of the products.

6.6 Competitor analysis

There are a number of competing systems. The paragraphs below compare those most relevant to our approach.

ViewScore - Very easy to navigate categories, thus the user can quickly find the highest ranking of products. It uses an intuitive scoring system with a 0 to 100 numerical score. One problem with this system is that it does not follow blogs that cover products; only the largest web site reviews are included. We think that including blogs covering products is important.

Wize - wize also uses a 0-100 scoring system, but the formula includes user comments and "reputation" as an integral part of the evaluation, as well as expert reviews. User reviews are very good, but we do not know if "reputation" can play a useful role. Is reputation based only upon the number of reviews or how valuable others have found these reviews? We do not care how many people talk about a new product, we are concerned about the how good the product. Ironically some of the products with high “reputation” have bad impression in off-line world. Wize’s product categories ranging from mobile phones, computers, to car seat.

Retrevo – Retrevo provides product specification which it gets from the manufacturer's web site. Retrevo also searches blogs and forums, professional reviews and articles, and provides an easy to switch between preview panes. It does not provide numerical ratings nor can the search results be retained. The variety of sources for products review is very good, but provides little added value.

6.7 Why should visitors trust reviews on my website

The reviews that we intent to collect are written by reviewers who were not paid to write reviews. These reviews are expected to mostly be written by customers who want to share their opinions, good or bad. All reviews have a URL so my site can provide a link to the original source, thus visitors can read the details that the original product review provides. This suggests that we need to perform some abstracting of the review - so that we can present a short view - but link to the original for the user who wants more information. There are text abstracting tools, see for example the thesis: http://www.csc.kth.se/~xmartin/papers/licthesis_xmartin_notrims.pdf

6.8 Business model

There are several alternative business models that could be used. Here we consider four of the potential business models.

1. Product review provider

Match a review to the partner’s directories through the internal identifier such as EAN, SKU, or ASIN, and then establish a feed that is updated daily. The whole process can be completed in a very short time. Review the information, and provide integration with the partner’s site(s).

2. Revenue sharing

Revenue sharing is also known as the cost of sales, for example my customers could pay a certain percentage of their sales revenue, which is based on the visitors' consumption due to a visitor to my website coming from my customers' websites.

3. Pay-per-click

With a pay-per-click approach, we would be paid by a customer web site based upon the number of clicks that their visitors make to my aggregated reviews.

4. Ads

Online Ads are a mature and common web business model. Ads on a vertical Business-to-business website have greater attractions to the companies in this industry, as compared to offline advertising the online ads. This approach could be used to target a particular user group.

7. Conclusions and Future work

This project focused on the Chinese e-commerce micro-environment, specifically how to use aggregated electronic product reviews in order to create a commerce website to attract more customers for both Business-to-business (B2B) and Business-to-customer (B2C) purposes. Some of the phases of this project that have already been implemented and some remain for the future work.

I achieved my goal of building a website and a database of aggregated reviews. I was able to collect reviews for products from existing web sites and aggregate them. However, at the beginning we doubted that we could implement this. I even thought of quiting due to my limited coding skills. Luckily I followed the advices of Professor Gerald Q. Maguire Jr. to finish what I am supposed to finish, not only in life in general, but especially in my professional activities.

I succeed in terms of the criteria stated in chapter 2 "This claim leads to one of our criteria for success, minimizing the amount of manually produced material -- as needing to manually produce material decreases the number of items that we can provide reviews for – hence potentially reducing our page rank. Minimizing the amount of manual work that is required is also important to minimize costs.".

One of the problems that remain is that a given item may have many different descriptions; as a result these appear as different products - rather than being aggregated into a single set of product reviews. Thus an item of future work is to improve the matching logic in order to match different product reviews from different sources through identifiers. It should be noted that in many cases the products might in fact be identical, but marked under different names - in order to target different customers or different markets; in this case more intelligent matching must be applied to decide when to aggregate and when not to aggregate the reviews. Specifically the amount of manual work that is now required to increase the number of reviews is simply adding new sites to the list of sites that the screen-scraper is to crawl.

Another task that needs to be done is to determine if these aggregated product reviews actually drive e-commerce buying decisions in the Chinese market.

References

[1] Sina science and technology,”Total number of China’s internet user up till 2.1billion” (2008-01-17), China, sina.com
[Webpage] http://tech.sina.com.cn/i/2008-01-17/13451980275.shtml
Last access on 2008-02-18

[2] Li Qi, “The status of China's Internet” (2000-11-27), China, cnw.com.cn
[Webpage]http://cnw2005.cnw.com.cn/issues/2000/46/4601.asp
Last access on 2008-02-21

[3] Luoheit Gov., http://www.luoheit.gov.cn/Article/ShowArticle.asp?ArticleID=1094
Last access on 2008-02-21

[4] Lipeng, “The People's Republic of China Regulations on Protection of Computer Information System Security”,1994
Last access on 2008-02-21

[5] CFCA, [Webpage] http://www.cfca.com.cn/
Last access on 2008-02-21

[6] Cnet.com, “CNET Reviews - Editors' Choice Reviews” (19 May 2009), USA, [Webpage]http://reviews.cnet.com/4924-5_7-0.xml?7eChoice=1&orderBy=-7rvDte&maxhits=25&dedup=1&tag=rb_content;rb_mtx
Last access on 2009-10-02

[7] Wikipedia, [Webpage] http://en.wikipedia.org/wiki/Data_mining
Last access on 2009-11-10

[8] Wen Jiabao, “Dissemination of the information network Protection Ordinance” (2006-05-29), China, State Council General Office
[Webpage] http://www.gov.cn/zwgk/2006-05/29/content_294000.htm
Last access on 2008-02-05

[9] eVOC Insights “Research retail”(March / April 2006), USA, evocinsights.com
[Webpage]http://www.evocinsights.com/research_retail.html
Last access on 2008-03-01

[10] PowerReviews, “White Paper: NetShops Case Study”, USA, publisher PowerReviews.com
[PDF]http://www.powerreviews.com/social-shopping/clients/netshops_case_study.pdf, Last access on 2008-03-02

[11] Network marketing management consultants,” Online store user reviews the role and preference survey” (2006-08-29), China, jingzhengli.cn
[Webpage] http://www.jingzhengli.cn/baogao/f20060829.htm
Last access on 2008-02-01

[12] Chris, “Promoting Your Book – The Power of Just One Recommendation”, 2008
[Webpage]
http://ckwebb.com/publishing/promoting-your-book-the-power-of-just-one-recommendation/
Last access on 2009-08-21

[13] Lauren freedman, “Merchant and customer perspectives on customer reviews and user-generated content” (February 2008), USA, powerreviews.com
[PDF]http://www.powerreviews.com/social-shopping/solutions/whitepaper/2008_WhitePaper_0204_4.pdf
Last access on 2008-03-02

[14] Bing.com,
[Webpage]http://www.bing.com/community/blogs/webmaster/archive/2008/06/03/robots-exclusion-protocol-joining-together-to-provide-better-documentation.aspxa
Last access on 2009-08-16

[15] Wikipedia.org,
[Webpage] http://en.wikipedia.org/wiki/Robots_exclusion_standard
Last access on 2009-08-16

[16] Robotstxt.org, [Webpage] http://www.robotstxt.org/
Last access on 2009-08-16

[17] Michael Schrenk, Webbots, Spiders, and Screen Scrapers:
A Guide to Developing Internet Agents with PHP/CURL. No Starch Press, 2007, 328 pages. ISBN 978-1-59327-120-6.
[Webpage]http://www.nostarch.com/frameset.php?startat=webbots

[18] Screen Scraper, [Webpage] http://www.screen-scraper.com/
Last access on 2009-11-10

[19] Alexa, USA
[Webpage]http://www.alexa.com/siteinfo/testfreaks.se
Last access on 2009-07-02

[20] Cal Henderson, Building Scalable Web Sites: Building, Scaling,
and Optimizing the Next Generation of Web Applications, O'Reilly Media, May 2006, 330 pages
ISBN-10: 0596102356 and ISBN -13: 978-0596102357

[21] Alipay, "Company Introduction", China, www.alipay.com
[Webpage]https://www.alipay.com/static/aboutalipay/about.htm
Last access on 2009-07-02

[22] China Internet Network Information Center
[Webpage]http://www.cnnic.net.cn/index.htm
Last access on 2009-11-10

Appendix A : Code for inserting products into the database

```
import java.sql.*;
String module = "Product to DB :";


void logger(String args) {

        //System.out.println("# "+args+" #");
        session.log(args);
}


Connection connect() {
        String con_url          = session.getVariable("DATABASE_CONNECT");

        // debug information
        if(con_url==null) {
                session.log(module+" : cannot save to database because variable DATABASE_CONNECT
has not been set");
        }

        ResultSet rs = null;
        Connection conn = null;

        try {
            Class.forName("sun.jdbc.odbc.JdbcOdbcDriver").newInstance();

        } catch(Exception e) {
                session.log(module+" : could not access the database driver !");
        }

        Connection conn = null;
        try {
        conn = DriverManager.getConnection(con_url);
        } catch(Exception e) {
                session.log(module+" error while trying to connect - "+e);
        }
```

```
    session.log(module+" : conn "+conn);
    return conn;
}

int resolveSource(Connection conn) {

        try {
            session.log(module+" resolveSource start");
            String query = "select * from Sources where SourceName = ? ";
            PreparedStatement prd = conn.prepareStatement(query);
            prd.setString(1,getSource());

            ResultSet r = prd.executeQuery();



            r.next();

            int value = r.getInt(1); // fetch the integer to retrieve

            session.log(module+" end of resolveSource "+value+" "+getSource());

            return value;

        } catch(Exception e) {
            logger(modulen+" "+e.toString());

        }

        return -1; // will never be written to db because it cant write to DBASE
}


void setVal(PreparedStatement prep,String what,int index,boolean w,boolean filter) {

        //session.log(module+"SetVal running!");
        String value = (session.getVariable(what) == null) ? "" : session.getVariable(what);
        //String value =    session.getVariable(what);


        if(filter) {
            session.log(module+"Using filter on "+value);
            value = cleanString(value);
            session.setVariable(what,value); // change the value for future scripts
        }

        //session.log(modulen+ " "+ what + " has value " + value);
        try {

            //prep.setObject(index,value.trim());
            if(value.trim().length()>0) {
                prep.setCharacterStream(index, new StringReader(value.trim()),
value.trim().length());
            } else {
                prep.setObject(index,value.trim());
            }

        } catch(Exception e) {
            logger(e.toString());
            e.printStackTrace();

        }

        if(w) {
                if(session.getVariable("NO_"+what+"_DELETE")==null) {
                    session.setVariable(what,null); // kill variable after entry
                }
        }
}

void setVal(PreparedStatement prep,String what,int index,boolean w) {
    setVal(prep,what,index,w,false);
}

void setVal(PreparedStatement prep,String what,int index) {
        setVal(prep,what,index,true,false);

}

// keep for the moment
void disconnect(Connection conn) {
    try {
                conn.close();
            } catch(Exception e) {


                                                        }
    session.log("connection to database closed!");
}
```

```
Connection start() {
    return connect();
}
void stop(Connection conn) {
    disconnect(conn);
}

String getSource() {

        return session.getVariable("SOURCE");
}

String fetchUrl() {
    // grab url to product
    // ref pages
    String turl = session.getVariable("TESTURL");
    String url;
    if(turl!=null && !turl.trim().equals("")) {
         url    = turl;
    } else {
      // if I dont manage to find the first product page, use the current page
         url = scrapeableFile.getCurrentURL();
   }
        session.setVariable("TESTURL",url); // set the url
        return url;
}



// does not count upwards, illegal to call before setID
//
// obsolete now
//
int getID() {

    int id_numb = Integer.valueOf(session.getVariable("_SESSION_ID"));
    return id_numb;
}



boolean shouldDBWrite() {
    String commitS = session.getVariable("DATABASE.COMMIT");
    session.setVariable("DATABASE.COMMIT",null);
```

```
        if(commitS!=null && (!commitS.equals("yes"))) {
                return false;
        }
        return true;
}

void setSource(PreparedStatement prep,Connection conn,int sv) {
                int SourceId = resolveSource(conn);
                String sid = (new Integer(SourceId)).toString();
                session.log(module+" preparing to setObject setSource!");
                prep.setObject(sv,sid); // set source id
}



        //
        //Remove things such as "MyProd&bsp;second" from a word and change it to be a string such as "MyProd second"
        //Should be moved to part that exists for both product and reviews
        //
String cleanString(String word) {
                //session.log(module+"Calling cleanString");
                if(word!=null) {

                                String textv = word;


textv = textv.replaceAll("&yacute;","�");




                                return textv;
                } else {
                        return null;
                }
}

/////////////////////////////// locals //////////////////////////////




        void insertProduct(Connection conn)    {

                session.log(module+" insertProduct");
                //
                String query = "insert into Products (Category, ProductName, PicURL,
Brand,SourceId,Testurl)"

     +"values (?,?,?,?,? ,?)";

        //  String vals = "insert into Products (CategoryName, ProductName, PicURL,
Brand,SourceId,Testurl) values ( 'agj','sad',10691,'mop' ,'ghty','jkl');";

                //System.out.println(vals);
                //Statement stm = conn.createStatement();
                //stm.executeUpdate(vals);


                PreparedStatement prep = conn.prepareStatement(query); // queries can be used to
retrieve automatic keys if needed


                // fetch product fields below



                //prep.setObject(3,sid); // set source id




                checkDuplicate(resolveSource(conn),conn);

                String product = session.getVariable("PRODUCT_NAME");
                if(product == NULL || product.trim().equals("")) {
                        throw new RuntimeException("Empty productName") ;
                }

                setVal(prep,"CATEGORY",1);
                setVal(prep,"PRODUCT_NAME",2,false,true);
                                setVal(prep,"PIC_URL",3);
                                        setVal(prep,"BRAND",4);
                //session.log(module+"Searching for source!");
                setSource(prep,conn,5);
                logger(module + "Source is "+sid);
                if(session.getVariable("TESTURL")==null) {
                        //session.setVariable("TESTURL",scrapeableFile.getCurrentURL());
                }

                fetchUrl();
                setVal(prep,"TESTURL",6);



                logger(modulen + " trying to execute query ");
                prep.executeUpdate();
                prep.close();
                // done
                logger("===================================================================
" + product + " commited to database
================================================================");
                try {
                        prep.close();


                } catch(Exception e) {
                        logger(modulen+ " error while closing prepStatement ");
                }


                session.log(module+" finished insertProduct");
        }


void insertIntoDB(String var,Connection conn) {

                try {
                if(session.getVariable(var)!=null) {



                        String query = "insert into Product_id (ProductName, Source_id, ID_kind, ID_value)
"
                                        +"values (? ,? ,? , ?) ";


                        PreparedStatement prep = conn.prepareStatement(query);
                        //session.log(module + " using setVal!");
                        setVal(prep,"PRODUCT_NAME",1,false);

                        //setVal(prep,"_SESSION_ID",1,false);
                        setSource(prep,conn,2);

                        //String sid = (new Integer(id)).toString();
                        //prep.setObject(1,sid);
                        //session.log(module + " id misc is " + id + " other " +
session.getVariable("_SESSION_ID"));

                        //setVal(prep,"SOURCE",2,false);
                        prep.setObject(3,var);
                        setVal(prep,var,4,true);

                        session.log(modulen + " using setVal from!");
                        session.log(modulen + " var="+var);

                        prep.executeUpdate();

                        // done
                        session.log("=== " + session.getVariable("PRODUCT_NAME")+ " " + var + "
value_id commited to database ==");
                        try {
                            prep.close();
                        } catch(Exception e) {
                            session.log(modulen+ " error while closing prepStatement ");
                        }

                }

                } catch(Exception e) {
                        if(session.getVariable("DB.ALWAYSPRODUCTID")==null) {
                            throw new Exception(e);
                        }
                }
        }


        String getDBMiscKey(int i) {
                return "DB.KEY"+(new Integer(i));
        }


        void insertMisc(Connection conn) {
                // insert MPN and UPC here with function
                if(session.getVariable("MPN")!=null && !session.getVariable("MPN").trim().equals("")) {
                        insertIntoDB("MPN",conn);
                }

                if(session.getVariable("UPC")!=null && !session.getVariable("UPC").trim().equals("") ) {
                        insertIntoDB("UPC",conn);
                }

                if(session.getVariable("PRODUCT_SUMMARY")!=null
&& !session.getVariable("PRODUCT_SUMMARY").trim().equals("") ) {
                        insertIntoDB("PRODUCT_SUMMARY",conn);
                }

                session.setVariable("NO_HOST_DELETE","1"); // standard value
                if(session.getVariable("HOST")!=null && !session.getVariable("HOST").trim().equals("") )
{
                        insertIntoDB("HOST",conn);
                }


                // below auto generated words
                int i = 0;
                while(session.getVariable(getDBMiscKey(i))!=null) {
                        //session.log("insertMisc number " + i + " -> " + getDBMiscKey(i));
                        insertIntoDB(session.getVariable(getDBMiscKey(i)),conn);
                        i++;
                }
        }



String getDBMiscKey(int i) {
        return "DB.KEY"+(new Integer(i));
}


void insertMiscValue(String val) {
        if(session.getVariable(val)!=null && !session.getVariable(val).trim().equals("") ) {
                        insertIntoDB(val,conn);
        }
}
```

```
void insertMisc(Connection conn) {

    // insert MPN and UPC here with function
    if(session.getVariable("MPN")!=null && !session.getVariable("MPN").trim().equals("")) {
        insertIntoDB("MPN",conn);
    }

    if(session.getVariable("UPC")!=null && !session.getVariable("UPC").trim().equals("") ) {
        insertIntoDB("UPC",conn);
    }

if(session.getVariable("EAN")!=null && !session.getVariable("EAN").trim().equals("") ) {
        insertIntoDB("EAN",conn);
    }

        if(session.getVariable("PRODUCT_SUMMARY")!=null
&& !session.getVariable("PRODUCT_SUMMARY").trim().equals("") ) {
        insertIntoDB("PRODUCT_SUMMARY",conn);
    }

    session.setVariable("NO_HOST_DELETE","1"); // standard value
    if(session.getVariable("HOST")!=null && !session.getVariable("HOST").trim().equals("") ) {
        insertIntoDB("HOST",conn);
    }


    // below auto generated words
    int i = 0;
    while(session.getVariable(getDBMiscKey(i))!=null) {
        //session.log("insertMisc number " + i + " -> " + getDBMiscKey(i));
        insertIntoDB(session.getVariable(getDBMiscKey(i)),conn);
        i++;
    }
}



    /**
     * Test for duplicated entries.
     */
boolean checkDuplicate(int SourceId,Connection conn) {

        String query = "select * from products where ProductName = ? and " +
            " SourceId = ? ";
        PreparedStatement prep = conn.prepareStatement(query);

        //session.log(modulen+" checkDuplicates running ");

        String pname = session.getVariable("PRODUCT_NAME");
        if(pname!=null) {
            pname = pname.trim();

            prep.setObject(1,pname);
            prep.setInt(2,SourceId);

            ResultSet res = prep.executeQuery();

            boolean result = res.next();

            // track a var
            if(result) {
                String var = session.getVariable("_duplicate_tracker");
                if(var==null || var.trim().equals("")) {
                    var = "0";
                }
                int val = Integer.parseInt(var)+1;
                session.setVariable("_duplicate_tracker",Integer.toString(val));
                session.log(modulen+" tracker found duplicate number "+val);
                return true;
            } else {
                // false and therefore I reset the duplicate number
                session.setVariable("_duplicate_tracker","0");
                return false;
            }
        }

        return false;
}

////////////////////////////////////////////////////////////////////


//session.log(module+" start");
Connection conn = start(); // always first

boolean skip = false;
```

```
String productname = session.getVariable("PRODUCT_NAME");
if(productname == null || productname.trim().equals("")) {
    skip = true;
    session.log(module+" warning: skipping product because productname is zero!");
}

session.log(module+" after start conn is "+ conn);
if(conn!=null && !skip) {

    //session.log("#conn !=null Passed ");

    try {
        session.log("#Shoulddbwrite!");
        if(shouldDBWrite()) {
            session.log(module+" setID");

            session.log(module+" conn");
            try {
                insertProduct(conn);
            } catch(Exception e) {
                if(session.getVariable("DB.ALWAYSPRODUCTID")==null) {
                    throw new Exception(e);
                }
            }

            // fix the rest later
            session.log(module+" insertMisc");
            insertMisc(conn);

        }
    } catch(Exception e) {
    session.setVariable("PRODUCT_NAME",null);
        session.log(module+" reporting "+e);
        session.log(module+" PRODUCT_NAME has been set to null as a response to the error,might be duplicate products");
    } finally {
            stop(conn); // last function in this command
    }
}
```

Appendix B : Code for inserting reviews into the database

```
import java.sql.*;
String module = "Review to DB :";
void logger(String args) {

    //System.out.println("# "+args+" #");
    session.log(args);
}

Connection connect() {
    String con_url          = session.getVariable("DATABASE_CONNECT");

    // debug information
    if(con_url==null) {
        session.log(module+" : cannot save to database because variable DATABASE_CONNECT
has not been set");
    }

    ResultSet rs = null;
    Connection conn = null;

  if(session.getVariable("__CONN")!=null) {
        conn = session.getVariable("__CONN");
  } else {
    try {
       Class.forName("sun.jdbc.odbc.JdbcOdbcDriver").newInstance();
        // this configuration string should be moved to a central configuration script
      } catch(Exception e) {
         session.log(module+" : could not access the database driver !");
      }

      //Connection conn = null;
      try {
     conn = DriverManager.getConnection(con_url);
      } catch(Exception e) {
         session.log(module+" error while trying to connect - "+e);
      }
        session.setVariable("__CONN",conn);
    }

        session.log(module+" : conn "+conn);
        return conn;
}

int resolveSource(Connection conn) {

                try {

                        String query = "select * from Sources where SourceId = ? ";
                        PreparedStatement prd = conn.prepareStatement(query);
                        prd.setString(1,getSource());

                        ResultSet r = prd.executeQuery();

                        r.next();

                        int value = r.getInt(1); // fetch the integer to retrieve

                        return value;

                } catch(Exception e) {
                        logger(modulen+" "+e.toString());

                }

                return 101; // will never be written to db because it cant write to DBASE
}

void setVal(PreparedStatement prep,String what,int index,boolean w,boolean filter) {

                //session.log(module+"SetVal running!");
                String value = (session.getVariable(what) == null) ? "" : session.getVariable(what);
                //String value =    session.getVariable(what);

            //filter = true; // always filter
                if(filter) {
                        session.log(module+" Using filter on "+what+"-->"+value);
                        value = cleanString(value);
                        session.setVariable(what,value); // change the value for future scripts
                }

                //session.log(modulen+ " "+ what + " has value " + value);
        try {

            //prep.setObject(index,value.trim());
            if(value.trim().length()>0) {
                prep.setCharacterStream(index, new StringReader(value.trim()),
value.trim().length());
            } else {
                prep.setObject(index,value.trim());
            }

        } catch(Exception e) {
            logger(e.toString());
            e.printStackTrace();

        }

        if(w) {
                if(session.getVariable("NO_"+what+"_DELETE")==null) {
                    session.setVariable(what,null); // kill variable after entry
                }
        }
}

void setVal(PreparedStatement prep,String what,int index,boolean w) {
//setVal(prep,what,index,w,true);
    setVal(prep,what,index,w,false);
}

void setVal(PreparedStatement prep,String what,int index) {
    setVal(prep,what,index,true,true);
}

// keep for the moment
void disconnect(Connection conn) {
    //try {
    //    conn.close();
    //} catch(Exception e) {

    //}
    //session.log("connection to database closed!");
}

Connection start() {
    return connect();
```

```
}
void stop(Connection conn) {
    disconnect(conn);
}

String getSource() {

        return session.getVariable("SOURCE");
}

String fetchUrl() {
    // grab url to product
    // ref pages
    String turl = session.getVariable("TESTURL");
    String url;
    if(turl!=null && !turl.trim().equals("")) {
         url    = turl;
    } else {
      // if I dont manage to find the first product page, use the current page
         url = scrapeableFile.getCurrentURL();
   }
        session.setVariable("TESTURL",url); // set the url
        return url;
}

//
int getID() {

    int id_numb = Integer.valueOf(session.getVariable("_SESSION_ID"));
    return id_numb;
}



boolean shouldDBWrite() {
    String commitS = session.getVariable("DATABASE.COMMIT");
    session.setVariable("DATABASE.COMMIT",null);
    if(commitS!=null && (!commitS.equals("yes"))) {
        return false;
    }
    return true;
}

void setSource(PreparedStatement prep,Connection conn,int sv) {
        int SourceId = resolveSource(conn);
        String sid = (new Integer(SourceId)).toString();
        prep.setObject(sv,sid); // set source id
}



    //
    //Remove things such as "MyProd&bsp;second" from a word and change it to be a string such as "MyProd second"
    //Should be moved to part that exists for both product and reviews
    //
String cleanString(String word) {
        //session.log(module+"Calling cleanString");
        if(word!=null) {

                String textv = word;



                return textv;
        } else {
            return null;
        }
}

//// locals ///

    void insertReview(Connection conn) {

        String query = "insert into Reviews (ProductName,Rating,Scale,ReviewDate,Good,Bad,Summary,Verdict,Author,Classify,Title,SourceId,Testurl)"
                        + "values (?,?,?,?,?,?,?,?,?,?,?,?,?) ";

PreparedStatement prep=null;
        if(session.getVariable("__REVIEWTODB")!=null) {
            prep = session.getVariable("__REVIEWTODB");
        } else {
            prep = conn.prepareStatement(query); // queries can be used to retrieve automatic keys if needed
        }

        setVal(prep,"PRODUCT_NAME",1,false);
        setVal(prep,"RATE",2);
        setVal(prep,"SCALE",3,false);
        setVal(prep,"DATE",4);
        setVal(prep,"PROS",5);
        setVal(prep,"CONS",6);
        setVal(prep,"SUMMARY",7);
        setVal(prep,"VERDICT",8);
        setVal(prep,"AUTHOR",9);

        if(session.getVariable("CLASSIFY")==null) {
            session.setVariable("CLASSIFY","PRO");
        }
        setVal(prep,"CLASSIFY",10);
        setVal(prep,"TITLE",11);

        // set source
        setSource(prep,conn,12);

        // set testurl
        fetchUrl();
        setVal(prep,"TESTURL",13);


        prep.executeUpdate();

        // done
        //session.log("=====================================================    ["    +
session.getVariable("PRODUCT_NAME")+"   |   "   +session.getVariable("AUTHOR")     +   "]   review
commited to database ==========================================");
      session.log("=====================================================Review
commited to reviews table ==========================================");
        try {
            prep.close();
        } catch(Exception e) {
            session.log(modulen+ " error while closing prepStatement ");
        }

    }


//session.log(module+" start");
Connection conn = start(); // always first
session.log(module+" after start conn is "+ conn);

boolean skip = false;
String productname = session.getVariable("PRODUCT_NAME");
if(productname == null || productname.trim().equals("")) {
    skip = true;
    session.log(module+" warning: skipping review because productname is zero!");
}

if(conn!=null && !skip) {

    try {
        if(shouldDBWrite()) {
            //session.log("insert Review");
            insertReview(conn);
            //session.log("finished review");
        }
    } catch(Exception e) {
        session.log(module+" reporting "+e);
    } finally {
            stop(conn); // last function in this command
    }
}
```

TRITA-ICT-EX-2009:208

www.kth.se