The CNET E-Commerce Data Set

Jennifer L. Beckham

University of Wisconsin, Madison
jbeckham@cs.wisc.edu

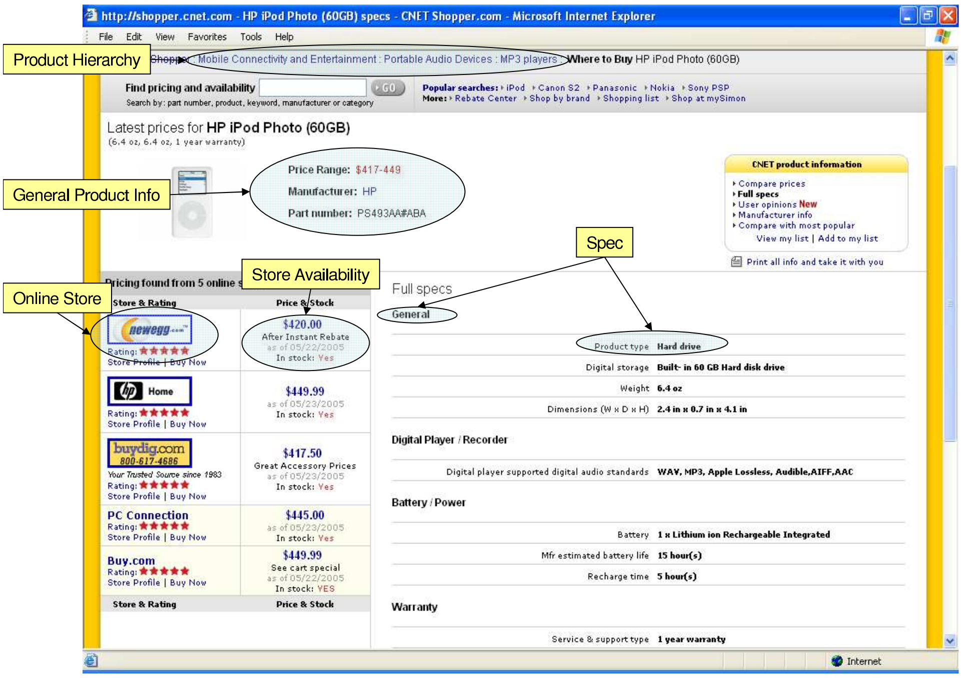
July 20, 2005

1 Introduction

E-commerce product catalogs often have product specifications (specs) that describe characteristics of each product. Specs are a list of attribute-value pairs where the attribute describes a property of the product and the value defines the property. Although there are numerous e-commerce catalogs online, there is very little known about the characteristics of specification data. Agrawal et al. [1] discuss a data set with 5,000 possible attributes and products only defining a few of the attributes, but discuss little else about the data itself. There are several characteristics of product specifications in e-commerce catalogs that are interesting from a database perspective. For example, the following questions help us to understand the specs within a catalog:

- How many products have specs?
- Of the products that define specs, how many products have values for the same attribute?
- Can products be grouped into categories based on the attributes that they define?

This document looks into answering these and other questions by describing the product specifications of a e-commerce catalog hosted by CNET Networks Inc. Shopper.CNET.com is a product catalog that provides specifications for a collection of technology products. With permission, we collected the product catalog from CNET over a one-week period in March



http://shopper.cnet.com - HP iPod Photo (60GB) specs - CNET Shopper.com - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Product Hierarchy
General Product Info
Online Store
Store Availability
Spec
Find pricing and availability
GO
Search by: part number, product, keyword, manufacturer or category
Popular searches: iPod Canon S2 Panasonic Nokia Sony PSP
More: Rebate Center Shop by brand Shopping list Shop at mySimon
CNET product information
Compare prices
Full specs
User opinions New
Manufacturer info
Compare with most popular
View my list | Add to my list
Print all info and take it with you
Internet

Shopper : Mobile Connectivity and Entertainment : Portable Audio Devices : MP3 players : Where to Buy HP iPod Photo (60GB)

Latest prices for **HP iPod Photo (60GB)**
(6.4 oz, 6.4 oz, 1 year warranty)

Price Range: $417-449
Manufacturer: HP
Part number: PS493AA#ABA

Pricing found from 5 online s

Store & Rating	Price & Stock
newegg.com Rating: ★★★★★ Store Profile | Buy Now	$420.00 After Instant Rebate as of 05/22/2005 In stock: Yes
hp Home Rating: ★★★★★ Store Profile | Buy Now	$449.99 as of 05/23/2005 In stock: Yes
buydig.com 800-617-4686 Your Trusted Source since 1983 Rating: ★★★★★ Store Profile | Buy Now	$417.50 Great Accessory Prices as of 05/23/2005 In stock: Yes
PC Connection Rating: ★★★★★ Store Profile | Buy Now	$445.00 as of 05/23/2005 In stock: Yes
Buy.com Rating: ★★★★★ Store Profile | Buy Now	$449.99 See cart special as of 05/22/2005 In stock: YES
Store & Rating	Price & Stock

Full specs

General

Product type	Hard drive
Digital storage	Built- in 60 GB Hard disk drive
Weight	6.4 oz
Dimensions (W x D x H)	2.4 in x 0.7 in x 4.1 in

Digital Player / Recorder

Digital player supported digital audio standards	WAV, MP3, Apple Lossless, Audible,AIFF,AAC

Battery / Power

Battery	1 x Lithium ion Rechargeable Integrated
Mfr estimated battery life	15 hour(s)
Recharge time	5 hour(s)

Warranty

Service & support type	1 year warranty

Figure 1: A product page for an Apple iPod on CNET.com.

2005. This document also describes our methods of data collection, data cleansing, and transformation from html to relational-style tables pairs.

2 Data Collection

We collected the CNET product catalog with permission from CNET in the seven days from March 30 to April 5, 2005. We obtained the data by first downloading the product catalog available at CNET's website [3]. Next, we retrieved, stored, and extracted the spec page for each product. Specs data on each html page are clearly marked as specs within the html along with the attributes and corresponding values.

Figure 1 shows a typical product page for a product on CNET.com. The figure highlights the information that we extracted from the pages: the product hierarchy, the general product information, product specs, and online stores. The information on the products ranges from all of the areas shown in the figure to only the product name. For example, some products

Attribute	Total	Ratio
Name	233297	1
Price Range	195299	0.84
Manufacturer	195658	0.84
Part Number	195582	0.84

Table 1: Primary attributes for products in the CNET data set and the number of occurrences.

do not have a part number and are not sold in online stores. The rest of this section discusses each extracted portion of the page.

2.1 Product Hierarchy

There are a total of 248,474 products and CNET assigned each product to a category within their product hierarchy. The product hierarchy has eight top-level categories, which in turn, have a total of 80 sub-categories. The eight primary categories and the number of products in them are *Computer Systems* (84,303), *Software* (70,038), *Office equipment* (24,125), *Mobile Connectivity and Entertainment* (22,012), *Games* (21,775), *Home Entertainment* (12,298), *Digital Photography and Video* (7,630), and *Tech Consumer Goods* (6,293).

The table in Appendix A lists the count of the products in each category and sub-category (labeled as *category:sub-category*). The third column, % with Specs, indicates the ratio of products with specs to total number of products in that sub-category. Some categories, such as Software, are well described by product specifications and over 85% have some product specifications, but other categories such as, Office equipment, are not described by product specifications at all.

2.2 General Product Information

Each product has core set of attributes that describe the general product information which includes the name, the price range (as defined by a range from low to high), the manufacturer, and the part number for the product. If nothing is known about the product, then the CNET site only lists the product name. Table 1 lists the four primary attributes for products in the data set and the number of products that have a value for that attribute. The table shows that all products in the data have a name and 84% of the products define a price range, manufacturer, or part number. When we

Statistic	Count
Number of unique attribute groups	208
Number of unique attribute names	1795
Number of unique attribute/group pairs	2984

Table 2: The count of the attributes and attribute groups.

discuss product specifications, we refer to the attributes of the spec that do not include name, price range, manufacturer, and part number.

2.3 Online Stores

Some products are sold on-line by vendors and these stores are indicated on each spec page. Vendors are described by their name and store rating. If a vendor sells a particular product, the page will include a price for the item as-of a certain day and an indication of how much stock is on-hand. Because we collected only the specs html page from the catalog, we only have the on-line vendor information that appears on the specs page, which is usually no more that 25 and on average 3 vendors per product. Of the data that we collected, there are 304 on-line stores that sell the products in the catalog. There are 195,355 (84%) products that are sold in at least one on-line store.

2.4 Product Specs

The product page has the specifications in the lower right-hand portion of the page, as illustrated in Figure 1, in a two-column table. The first column contains an attribute name and the second column contains the value for the attribute. Also in the first column, in bold, attributes are grouped into a higher-level categories, such as General and Battery/Power as shown in the figure. There are 1795 distinct attribute names, 208 high-level attribute groups, and 2984 group/attribute name parings. We consider the group/attribute name parings as the true *attributes* of the spec because the groups distinguish among attributes with the same name. For instance, for a product that has a screen (like an iPod), dimension attributes of height and width can describe the entire product as well as the screen size of the product. In cases like dimension, a product will define the same attribute, but place it under a different category.

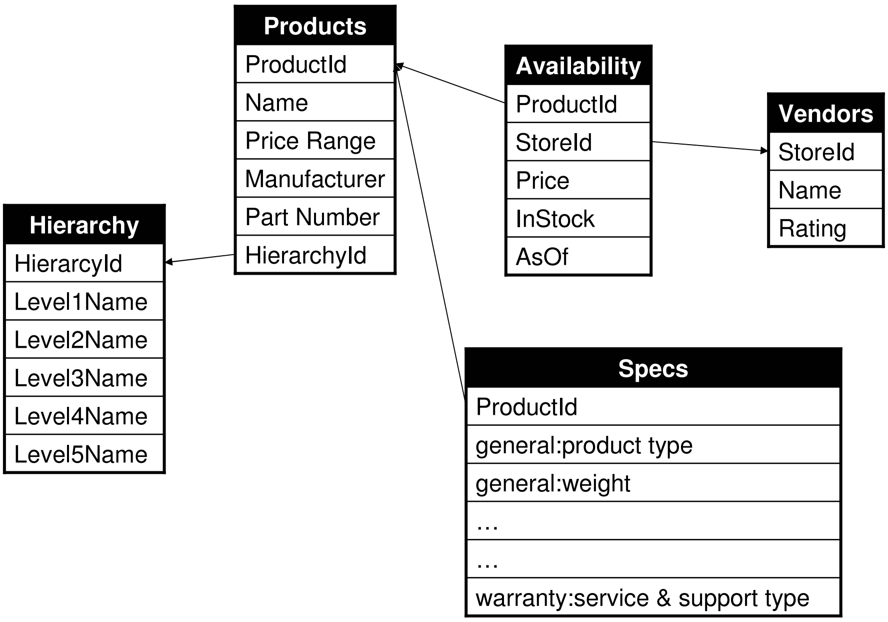


Figure 2: The schema used to store the product information.

3 Relational Transformations and Tables

The data found on the html specs pages at CNET.com was extracted and transformed into relational tables. The transformation to relational tables was done with as little change, or cleaning, to the actual data on the web pages as possible. We keep the data in its natural state because it best represents what is available currently in product specs and how companies describe their products.

Figure 2 describes the schema used for the data set. The data is split into tables Products, Hierarchy, Specs, Availability, and Vendors. The data is imperfect and there are many opportunities for data cleaning and domain normalization. We did basic cleaning on the data and changed all of the meta-data and data to lower-case. The lower-case transformation allows attributes and groups that are the same name, but with different case, to be mapped the same. We also assure that each attribute/group pair appears at most once per product. If there are attribute/group pairs with the same name within a specs, the values are concatenated and combined with a semicolon. There are 2464 occurrences of the same attribute/group pair appearing more than once (in fact, none appear more than twice) in the same product specification.

In determining domains for the attribute/group pairs, we used simple pattern matching to find attributes that have only numbers as the values. There are 172 attribute/group pairs with only numbers as the domain. The data has several other numeric domains that are labeled as a measurement (such as in inches), but different products describe the attribute with as multiple metrics such as inches or centimeters. The extraction of these data types and transformations are beyond the scope of this paper.

The specs table can be stored in two alternative table representations. One representation maps the attribute/groups to columns of a table and cells within the relation contain the values for the attributes. The alternate representation stores the specs as a vertical table with each spec represented as a set of product id, attribute, group, and value tuples. The vertical alternate is similar to what is described in Agrawal et. al. [1].

4 Data Statistics and Distributions

This section discusses the data statistics and distribution for the attributes and products in the data set. We look at two aspects of the data by focusing on the per attribute statistics and the per product statistics. Conceptually,

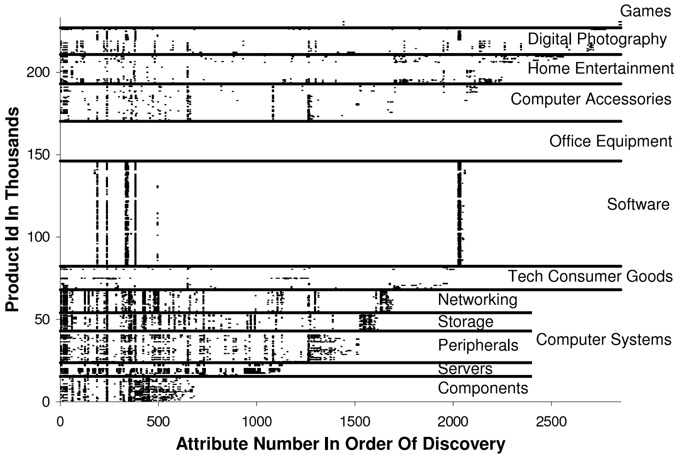


Figure 3: A sample of the CNET spec data.

attribute statistics explores the column-wise properties of the specs and the product statistics looks at the row-wise properties of the specs.

Within each discussion we breakdown the statistics according to the product hierarchy. Figure 3 shows a sample of the data. The figure is organized as it was collected from the product catalog. The attributes are numbed along the x-axis as they were discovered in the data and the y-axis numbers products in alphabetical order within each product category. The horizontal lines separate the categories in the product hierarchy and breakdown of the some of the larger sub-categories in Computer Systems. The ordering of the attributes as they were discovered in the data allows us to see some of the correlation between attributes and the product categories that use them. There is a lot of overlap in attributes across the categories, which indicates that there are connections between domains. There are some attributes that get heavily used within a category and not used at all within other categories, which indicates that some attributes are relatively specific to the products in the categories that they are defined.

4.1 Product-wise

Specifications for each product vary in the exact attributes that get defined and the number of attributes per product that get defined. Over the entire data set, the average product defines eleven attributes and the mode number of attributes defined is five. Figure 4 shows the distribution of the number of attributes defined by products in the catalog. A majority of products

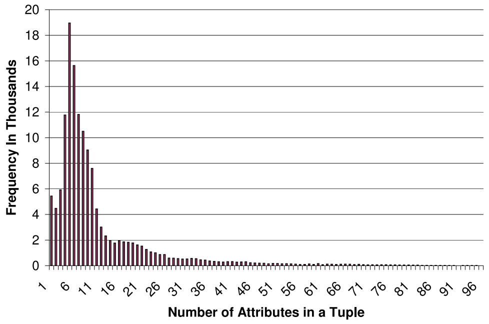


Figure 4: The frequency of the number of attributes in a tuple.

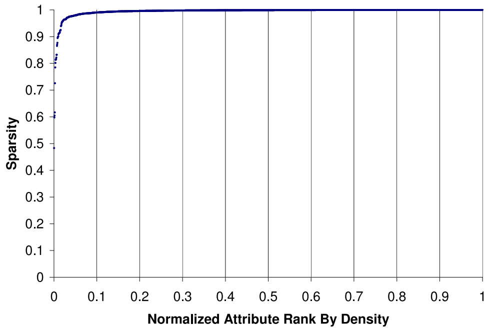


Figure 5: The sparsity of attributes within the data. The attributes are normalized to the total number of attributes and ordered according to the degree of sparsity.

(71%) have ten or fewer attributes, but there are some products that have many attributes.

Table 4 in Appendix A lists the average number of attributes for products with specs in each category in the product hierarchy. The table shows that some categories describe products with many attributes, but other categories do not have as detailed specifications.

4.2 Attribute-wise

This section looks at the properties of individual attributes within the specs. Figure 5 orders the sparsity of the attributes over their normalized ranks. The figure shows that 98% of the attributes are more than 95% sparse, and roughly follows a Zipf distribution where the frequencies are inversely proportional to their ranks. Cheng et. al. [2] made a similar observation about attributes in structured e-commerce databases on the web.

Group\Attribute	Sparsity %
software\license qty	24%
general\compatibility	50
software\license type	60
software\license pricing	60
warranty\service & support type	60
system requirements\OS required	72
expansion port(s) required\slot(s) required	74
service and support\service and support details	81
general\product type	82
system requirements\min processor type	83

Table 3: The top ten least sparse (most dense) attributes and their sparsity within the specs.

Table 3 lists the top ten least sparse attributes and their sparsity within the specs table. Most of the attributes appear in in products within the hierarchical category Software.

5 Summary

The CNET data set is a representative e-commerce data set that has several desired properties for using it as a benchmark for e-commerce search. The products are organized into hierarchies, are sold by online stores, have user ratings, and each product can have a specification that describes the properties. We have downloaded and extracted the information from CNET and have described a relational schema for storage and querying of the data. The specifications are relatively sparse and on average each product only defines a handful of attributes.

References

[1] R. Agrawal, A. Somani, and Y. Xu. Storage and querying of e-commerce data. In *VLDB*, pages 149–158, 2001.

[2] Kevin Chen-Chuan Chang, Bin He, Chengkai Li, Mitesh Patel, and Zhen Zhang. Structured databases on the web: observations and implications. *SIGMOD Rec.*, 33(3):61–70, 2004.

[3] CNET Product Directory. `http://shopper.cnet.com/4296-3000_9-0-0-0.html`, March 2005.

A Appendix

Table 4: Product Categories and the percentage of products within each category that have specifications. The third column shows the average number of attributes for the category when the product has specs.

Category Name	Products	% w/Specs	Ave Attrs
Computer Systems	9534	0%	0
:Desktops	646	87	50
:Graphics & Sound	1497	90	13
:SOHO Servers	1683	90	26
:Notebooks	1992	89	65
:Storage	9879	93	16
:Networking	12960	93	13
:Components	13910	92	12
:Computer Systems Accessories	15951	74	4
:Peripherals	16251	80	12
Total Computer Systems	84303	76	23
Software	7668	0%	0
:Handheld software	173	94	6
:Home and personal	233	77	9
:Music and video	768	96	8
:Education and reference	1567	94	7
:Operating systems	2187	73	2
:Development tools	2638	96	6
:Graphics and publishing	2656	97	8
:Business & productivity	4205	97	7
:System utilities	4970	99	7
:Networking tools	5401	96	5
:Internet utilities	7163	99	6
:Games	15177	95	6
:Data management	15232	95	6
Total Software	70038	85	6
Office equipment	3447	0%	0
:Calendars and planners	150	0	0
:Business cases	177	0	0
:Custom imprints	204	0	0
:Executive gifts	303	0	0
:Desk accessories	811	0	0
:Pens, pencils, and markers	902	0	0
:Maintenance and breakroom	1144	0	0
:A/V supplies and equipment	1365	0	0
:Office machines	1672	0	0
:Paper forms and envelopes	1676	0	0

Category Name	Products	% w/Specs	Ave Attrs
:Filing binders and storage	3807	0	0
:Basic supplies and labels	4091	0	0
:Furniture	4376	0	0
Total Office equipment	24125	0	0
Mobile Connectivity and Entertainment	6464	0%	0
:Portable Video Devices	91	96	29
:Tablet PCs	153	100	36
:Handheld Devices	283	88	25
:Handheld Accessories	958	14	1
:Portable Audio Devices	3141	64	20
:Personal Comm. and Navigation	4486	26	9
:Portable Electronics Accessories	6436	40	6
Total Mobile Connectivity and Entertainment	22012	29	18
Games	4369	0%	0
:DS	14	0	0
:Consoles	16	88	23
:PSP	21	0	0
:Gamecube	209	12	2
:PC	224	80	7
:Game Accessories	227	83	7
:Game Boy Advance	235	0.04	2
:XBox	253	14	3
:Playstation2	478	23	5
:Legacy game platforms	552	21	4
Total Games	6598	10	6
Home Entertainment	2068	0%	0
:TV/HDTV Tuners & Receivers	16	94	18
:Video Game Consoles	16	88	23
:Digital Media Receivers	43	88	19
:Stereo & Home Theater Systems	307	84	42
:Video Players and Recorders	518	86	28
:Audio System Components	694	78	35
:Speakers & Speaker Systems	922	75	16
:TVs	1134	83	39
:Home Entertainment Accessories	6580	47	10
Total Home Entertainment	12298	49	25
Digital Photo. and Video	1885	0%	0
:Photo Printers	32	75	32
:Digital Camcorders	238	81	45
:WebCams	249	52	13
:Digital Cameras	499	91	59
:Digital Camcorder Accessories	873	25	1
:Digital Camera Accessories	3854	52	8

Category Name	Products	% w/Specs	Ave Attrs
Total Digital Photography and Video	7630	39	26
Tech Consumer Goods	823	0%	0
:Cameras (non-digital)	312	76	36
:Fixed location telephony	709	66	33
:Specialized electronics	1861	89	6
:Car electronics	2588	36	17
Total Tech Consumer Goods	6293	52	23
Total Products	233304	61	11