Automatic Organization of Product Information by Using Product Categories

Tadanori Aruga

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 2000

Keywords: World-Wide Web, Product List, Product Information, Information Extraction, Automated editing.

A lot of product information came to be on World Wide Web (WWW) by the spread of the Internet. There is already very useful information for shopping on WWW.

There are two phases in the decision-making at shopping: product selection and shop selection.

In the product selection, which is the first phase, we determine which product to purchase. At the time we start shopping, we have already decided what kind of product to purchase. For example, he has already decided that he buys television. *The kind of product* is called *product category*. The product selection phase is to select the exact product to purchase in a certain product category.

In the shop selection, which is the second phase, we determine where we purchase. Many shops sell many products; one kind of product may be sold at two or more shops at different prices. Therefore, we compare prices at several shops, and choose the best shop.

In both of two phases of shopping, *comparison* is the key issue. However, the current WWW does not provide enough help for such comparison.

The comparison in the product selection phase, we have to obtain information about two or more products in a specific product category. However, such information exits separately on different web sites of product manufacturers. For example, when we want to purchase a television, we must visit two or more web sites of different manufactures, and check features of every television on these web sites. We have to take almost the same trouble at the shop selection phase.

The reason why such trouble happens is that WWW is not organized to the convenience of the users.

I surmise that the most convenient organization for the users is as follows.

1. At the product selection phase.
   Product information should be organized for every product category so that we can overview information of all products in a certain product category at once without manufacturer's boundary.

2. At shop selection phase.
   Price information should be organized for every product so that we can overview prices of all shops at once without shop's boundary.

On the current WWW, information is organized as follows.

1. Product information is organized by every manufacturer in its web site.

2. Price information is organized by every shop in its web site.

The gap between the current organization and the desirable organization can be fulfilled by reorganization of product information.

There have been several researches in reorganization of the product information to support shop selection, but there is no research in reorganization of the product information to support product selection.

In this paper, I propose a system that reorganizes product information: it extracts product information from each manufacturer's web site, and generates a product list for each product category. The product list contains all products in the category without manufacturer's boundary; for each product, it provides a set of important attributes.

The system consists of five modules: collection module, price investigation module, product-information database, user interface, and domain knowledge.

The collection module accepts a URL that is the root URL of a manufacturer's web site, and extracts product information on the web site. The module consists of two sub modules: category-URL collection and information extraction. Usually, in the manufacturer's web site, there is the table-of-contents of the products that the manufacturer produces; the table-of-contents contains a set of links called as category-URLs. Each category-URL provides the entrance to all products in the category. The category-URL collection module tries to find such kind of table-of-contents and extracts a set of category-URLs.

The information extraction module accepts a category-URL and extract product information of all products in the category. This module first collects the pages that provide product information by using two methods. The first method uses the product list page that contains links to all product pages in the category. The second method uses the depth-bounded blind search and selection by using several features of product pages.

¿From collected pages, the information extraction sub module extracts product information by three methods: table analysis, heading analysis, and enhanced-sting extraction. All extracted information is stored into the product-information database.

The price investigation module investigates the selling price of each product by checking on-line shop sites. The obtained prices are also stored into the product-information database.

The user interface accepts a product category as a user's request and generates the product list of the category.

The system is designed for the home electronics products: the knowledge that is peculiar to this domain is written in the domain knowledge.

I conducted the experiment on product information collection: the system collects product information from five manufacturers. The system collected all twenty two table-of-contents that contain category-URLs. The system collected information about eighty four products out of ninety five products on the web sites: the recall was 88 percent. The system succeeded to extract a set of important attributes for seventy four products out of ninety five products: the precision was 77 percent. The result is satisfactory in spite of the various layouts of manufacturers' web sites.