Automated Dictionary Discovery for the Online Marketplace

Fei Chiang
Dept. of Computer Science
University of Toronto
Toronto, Canada
fchiang@cs.toronto.edu

Renée J. Miller
Dept. of Computer Science
University of Toronto
Toronto, Canada
miller@cs.toronto.edu

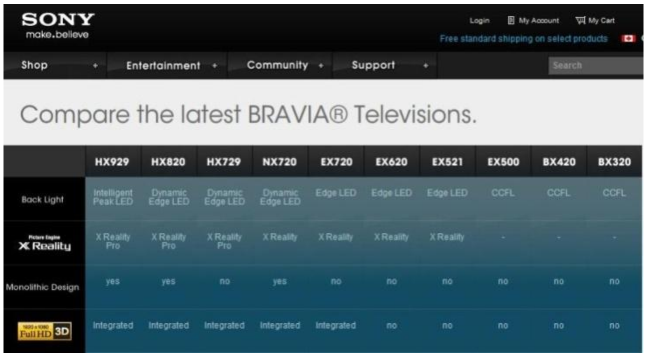
ABSTRACT

Shopping online has become a prolific activity as the number of online vendors and consumers continue to rise each year. In 2009, almost $15 billion in goods and services were ordered online by Canadians [1]. About 53% of these consumers 'window shop' by doing product research before actually making a purchase. Therefore, it is important that online vendors provide up-to-date and accurate product information to assist users in making educated decisions. In this poster, we present a tool that discovers product features, which will assist vendors and consumers to more accurately compare products in the online marketplace

1. INTRODUCTION

Online shopping has become a prolific activity in the last decade. Consumers can purchase almost any product or service imaginable as more vendors have realized the opportunity to increase sales and promote advertising via the online marketplace. For consumers, the main benefit of online shopping is the convenience of purchasing products without having to visit an in-store retailer. Consequently, online shoppers normally have an intended product, and are more knowledgeable about their purchase. Since there is often no sales associate available to answer questions online, consumers rely on the product descriptions, reviews, and product comparison features to make educated purchase decisions. To accurately compare a set of similar products, vendors and shoppers need to identify the discriminating features (attributes) and options of each item. For example, comparing two smoke detectors may involve looking at the size, power supply options, and whether it supports carbon monoxide detection. Each of these features can be represented with the attributes: *Size, Power, AdvancedOptions*. Having an online system automatically identify these discriminating attributes and the values for each attribute is a difficult task. There are many manufacturers producing similar products in a given domain. Furthermore, to identify these attribute values across different domains, ranging from laptops to children's clothing, is a challenging task.

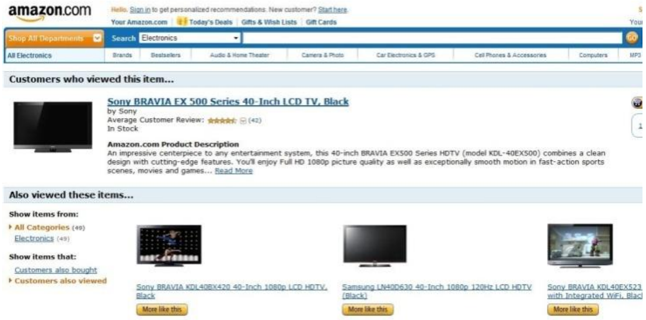
iConference 2012 February 7-10, 2012, Toronto, Ontario, Canada
Copyright 2012 ACM 978-1-4503-0782-6…$10.00.



Compare the latest BRAVIA® Televisions.

	HX929	HX820	HX729	NX720	EX720	EX620	EX521	EX500	BX420	BX320
Back Light	Intelligent Peak LED	Dynamic Edge LED	Dynamic Edge LED	Dynamic Edge LED	Edge LED	Edge LED	Edge LED	CCFL	CCFL	CCFL
Picture Engine X Reality	X Reality Pro	X Reality Pro	X Reality Pro	X Reality	X Reality	X Reality	X Reality	-	-	-
Monolithic Design	yes	yes	no	yes	no	no	no	no	no	no
1920 x 1080 Full HD 3D	Integrated	Integrated	Integrated	Integrated	Integrated	no	no	no	no	no

Figure 1: TV models product comparison



Customers who viewed this item...

Sony BRAVIA EX 500 Series 40-Inch LCD TV, Black
by Sony
Average Customer Review: (42)
In Stock

Amazon.com Product Description
An impressive centerpiece to any entertainment system, this 40-inch BRAVIA EX500 Series HDTV (model KDL-40EX500) combines a clean design with cutting-edge features. You'll enjoy Full HD 1080p picture quality as well as exceptionally smooth motion in fast-action sports scenes, movies and games... Read More

Also viewed these items...

Show items from:
All Categories (49)
Electronics (49)

Show items that:
Customers also bought
Customers also viewed

Sony BRAVIA KDL40BX420 40-Inch 1080p LCD HDTV, Black
More like this

Samsung LN40D630 40-Inch 1080p 120Hz LCD HDTV (Black)
More like this

Sony BRAVIA KDL40EX523 with Integrated WiFi, Blac
More like this

Figure 2: TV product description

Some websites, such as store.sony.ca, allow users to compare product models from the same manufacturer based on a set of attributes. An example comparison of Sony TV models is shown in Figure 1. This comparison is a relatively straightforward task since Sony is able to define the attributes being compared, can recognize valid attribute values, and has customized its website towards this comparison. If we consider a broader online retailer such as amazon.com, the product comparison feature across different manufacturers and models is not as straightforward, and is currently not available. Users are simply shown the product description page, and some similar products (based on customers' viewing history), as shown in Figure 2. Users must manually determine the desired attributes, and toggle between products to do the comparison, which can be tedious and time consuming. Online systems are unable to recognize the key attributes of a general product, and are unable to determine the valid attributes values to do a comparison. This list of valid attribute values is called a *dictionary*. For example, a possible dictionary for TV manufacturers would consist of {Sony, Samsung, Panasonic, JVC}. In this poster, we present *AutoDict*, a tool for automated

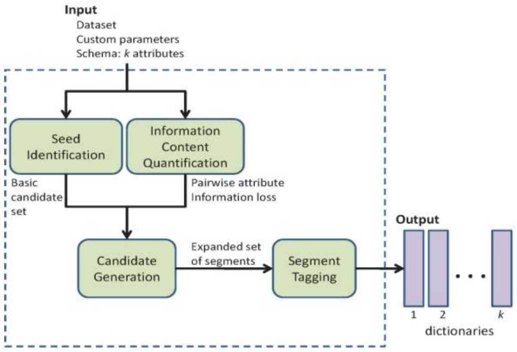


Figure 3: Dictionary discovery framework

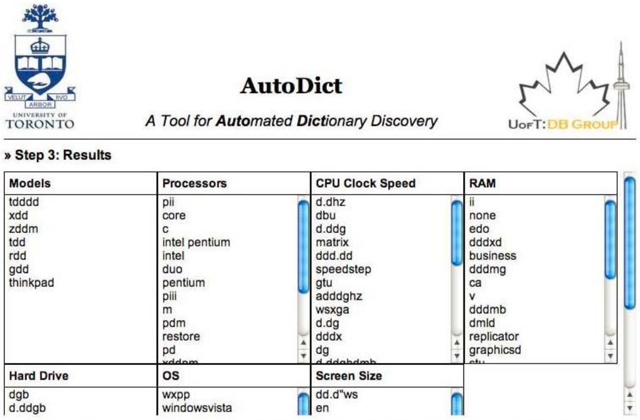
dictionary discovery that helps online systems and users identify the relevant features of a product, and produces a dictionary containing valid values for a given attribute feature.

Previous techniques have applied machine learning models that reference dictionaries to identify the attributes [2, 4]. Work in online query segmentation [6] and keyword tagging [5] have focused on using generative models to maximize the probability that a candidate segmentation is the correct one. A semi-automatic dictionary discovery tool is proposed by Godbole et al. [3] that assumes the attributes are given, but is unable to handle out of domain values correctly. Most prior work has not addressed the question as to how these dictionaries are obtained. Generating and customizing such a list of values for an application or online vendor is both time consuming and requires highly specific domain knowledge. In this poster, we present the details and usage of our dictionary discovery tool, that will build dictionaries to enable vendors and users better understand their information.

2. FRAMEWORK

Figure 3 illustrates the architecture of AutoDict consisting of modules for identifying and expanding segments from input data, and tagging them to the appropriate attribute. The system takes a data file, a set of user specified threshold parameters, and a set of *k* attributes as input, and produces *k* populated dictionaries according to these attributes. The first phase, *seed identification*, consists of identifying all the frequent sets of words occurring in the dataset which we call *seeds*. These seeds are ranked and refined according to the amount of information they capture, which is determined in the *information content quantification* module. At this point, we have a basic set of candidate segments. Since the quality of a dictionary is subjective according to the application needs, we allow the user to associate particular segments to specific dictionaries to produce more relevant dictionaries. We expand upon this basic set of candidates in the *candidate generation* module by considering candidate values that may be composed of non-sequential words in the data. Finally, in the *segment tagging* component, for each product description in the dataset, we tag each segment in a record with one of the *k* dictionary labels, thereby producing *k* populated dictionaries.

Online search engines and shopping sites can apply these dictionaries to compare products, by checking that a product feature is valid and refers to the same attribute. By having a correct set of attribute values, online vendors can provide faster and more relevant search results to correctly identify products, thereby providing a better shopping experience to the consumer. AutoDict's graphical display is shown in Figure 4, displaying a set of dictionaries in the laptops domain.



AutoDict

A Tool for ***Auto****mated* ***Dict****ionary Discovery*

» Step 3: Results

Models	Processors	CPU Clock Speed	RAM
tdddd	pii	d.dhz	ii
xdd	core	dbu	none
zddm	c	d.ddg	edo
tdd	intel pentium	matrix	dddxd
rdd	intel	ddd.dd	business
gdd	duo	speedstep	dddmg
thinkpad	pentium	gtu	ca
	piii	adddghz	v
	m	wsxga	dddmb
	pdm	d.dg	dmld
	restore	dddx	replicator
	pd	dg	graphicsd

Hard Drive	OS	Screen Size
dgb	wxpp	dd.d"ws
d.ddgb	windowsvista	en

Figure 4: Populated dictionaries

3. PRELIMINARY EVALUATION

We evaluated the quality of the discovered dictionaries, and the initial results are promising. We evaluated our tool using three real datasets from the address, laptops, and bicycles domain. The initial precision and recall results from the address domain are both greater than 98%, highlighting the quality of the discovered attribute values. We plan to do more extensive quality tests with the laptops and bicycles data, and to incorporate user feedback to re-generate dictionaries based on user input.

4. CONCLUDING REMARKS

We have presented a framework for automatically discovering attribute dictionaries. By having a complete and accurate set of dictionaries, online vendors will be able to provide more detailed product comparisons and return more relevant search results, both enriching the consumer shopping experience. Our AutoDict tool has handled the process of detecting and extracting these attribute values, thereby saving time and increasing efficiency. As our next steps, we plan to conduct more extensive tests using data from different domain areas, and to consider how user preferences can be incorporated into the framework.

5. ACKNOWLEDGMENTS

This work was partially supported by the NSERC Business Intelligence Network.

6. REFERENCES

[1] Statistics Canada Report: www.statcan.gc.ca/dailyquotidien/ 100927/dq100927a-eng.htm.

[2] V. Borkar, K. Deshmukh, and S. Sarawagi. Automatic segmentation of text into structured records. *SIGMOD Rec.*, 30(2), 2001.

[3] S. Godbole, I. Bhattacharya, A. Gupta, and A. Verma. Building re-usable dictionary repositories for real-world text mining. In *CIKM*, pages 1189–1198, 2010.

[4] S. Sarawagi and W. W. Cohen. Semi-markov conditional random fields for information extraction. In *NIPS*, 2004.

[5] N. Sarkas, S. Paparizos, and P. Tsaparas. Structured annotations of web queries. In SIGMOD Conference, pages 771–782, 2010.

[6] B. Tan and F. Peng. Unsupervised query segmentation using generative language models and wikipedia. In *WWW*, pages 347–356, 2008.