Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2014 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

2014

A NOVEL FRAMEWORK BASED ON WORD-OF-MOUTH MINING FOR NON-PROSUMER DECISION SUPPORT

Heng Tang
City University of Hong Kong, hengtang@umac.mo

A NOVEL FRAMEWORK BASED ON WORD-OF-MOUTH MINING FOR NON-PROSUMER DECISION SUPPORT

Heng Tang, Faculty of Business Administration, University of Macao, Macao, China, hengtang@umac.mo

Abstract

The deeper penetration of business-to-consumer e-commerce requires that customer decision support systems (CDSS) serve a wider range of users. However, a significant weakness of existing e-shopping assistance programs is their inability to aid non-professional consumers (non-prosumers) in buying highly differentiated products. This paper proposes a novel framework that infers product recommendations with minimal information input. At the heart of the proposed framework is the feature-usage map (FUM), a Bayesian network-based model that encodes the correlations among a product's technical specifications and its suitability in terms of its using scenario (usage). It also incorporates a query-based lazy learning mechanism that elicits a product's rating score from product reviews and constructs its corresponding FUM in an on-demand manner. This mechanism allows the knowledge base to be enriched incrementally, with no need for an exhaustive repository of FUMs pertaining to all possible usage queries a user may invoke. The effectiveness of the proposed framework is evaluated through an empirical user study. The results show that the framework is able to effectively derive product ratings based on specified usage. Moreover, this rating information can also be incorporated into a conventional buying guide system to deliver purchase decision support for non-prosumers.

Keywords: Customer Decision Support, Natural Language Processing, Opinion Mining, Bayesian Networks.

1 INTRODUCTION

Sophie wants to buy a new digital camera from Amazon for her graduation trip to Hawaii, during which she will spend much of her time scuba diving. She is looking for a portable camera capable of capturing casual shots on the beach or underwater. Sophie is not tech-savvy, and knows little about either photography or electronic products in general, and she thus googles “digital camera for diving” and browses the results. The first hit is a list of cameras of a particular brand, followed by a number of webpages offering tips for taking photos underwater and using terms that are Greek to her. Sophie then visits a professional digital camera review website, Dpreview.com[1], which provides a purchase assisting tool called “buying guide” that helps users to single out the most appropriate camera from the site’s camera database. However, she is totally lost when she is asked to choose such features as “aperture,” “max ISO,” “prime lens,” “sensor size,” “exposure bracketing,” and the like. After an hour of frustrating searching and browsing, Sophie gives up and approaches to her local Best Buy store, hoping that she can get some advice from a real shop assistant.

Product differentiation strategy has been widely used by nowadays manufactures in order to reach diverse segments of the market (Kotler & Keller, 2006). For example, most consumer electronics (e.g., digital cameras, smartphones, and personal computers), automobiles, and household appliances are differentiated not only vertically but also horizontally so as to attract various customer groups. To this end, products are designed with various configurations of technical specifications[2]. For instance, a single digital camera model can be characterized by over 20 specialized specification items, including sensor size, resolution, effective pixels, phase detect focus, and constant aperture, and an automobile has even more in categories ranging from chassis and engine to fuel effectiveness. As a result, only a very small proportion of buyers, those termed “prosumers,” are equipped with adequate domain knowledge to study product specifications, read reviews, and use conventional feature-based purchasing guides to spot suitable products. For most normal consumers such as Sophie, in contrast, information overload and the use of obscure jargon have become major barriers to making purchase decisions with the aid of feature-based purchasing assistants.

This research designs a framework aiming to assist non-prosumers to select products most suitable for a specific using scenario (referred to as “usage” herein). Such a premise is based on the observation that the first step in typical buying behavior is to identify the need (Kotler & Keller, 2006), in other words, what the consumer set out to do with the product. For instance, a compact water-resistance camera could be more suitable than an expensive high-end professional camera in many typical using scenarios – the latter would be too heavy and bulky for outdoor activities or lack water-resistance for safe use at the beach. This starting point coincides with the common scenario that a purchasing guide in a brick-and-mortar store usually asks a novice buyer "What do you buy this camera/car/computer for?” A plausible method to identify a suitable product for an intended usage would be to summarize the word-of-mouth on the Internet (eWoM) surrounding those products with regard to the given usage, as products that have gained a high degree of customer satisfaction normally enjoy positive eWoM (Chen, 2011). Hence, Opinion mining, a technique that exploits natural language processing (NLP), computational linguistics, and text analytics, is making inroads in this arena (Scaffidi, et al., 2007). In studies of opinion mining, products’ comment polarity is derived through analysis of their review articles. Accordingly, products with a high overall polarity score can be regarded as positively evaluated by reviewers, and thus suitable for recommendation to customers. Such review summarization techniques offer a viable solution to the aforementioned recommendation problem, that is, the system retrieves from the review corpus a key term or short phrase describing the using scenario (e.g., travel or scuba diving) and locates the sentence(s) containing it. Thus, a product’s overall appraisal score with regard to a given usage can be viewed as the average polarity point. A straightforward calculation of this appraisal score would be $score_{u,p} = average(polarity(s_{u,p}))$, where s_i is a review sentence about product p containing usage term u. Products with high appraisal scores can be recommended to customers.

[1] http://www.dpreview.com, retrieved in March 2013

[2] In this paper, the two terms “specification” and “feature” are used interchangeably.

However, the review summarization approach is not always practical in real-world applications because only popular product models attract numerous online comments. As a result, niche products are ignored by the system owing to the scarcity of reviews. Worse still, the approach fails entirely in recommending newly released models because early reviews for new product models are rare. CDSS based on straightforward review summarization thus suffer from limited recommendation coverage and bias.

To overcome this problem, the research reported herein incorporated review summarization and rating derivation via product features. It is recognized that the features possessed by a product play the most important role in its performance and usability in certain using scenarios. For instance, a lightweight digital camera with a wide zoom range can be assumed to be a good camera for travel. Accordingly, “Weight” and “Zoom Range” are product features, and “travel” a usage term, of digital cameras. This paper proposes a framework for summarizing product reviews and deriving the underlying dependencies between product features and using scenarios. A probabilistic graphical model, the Feature-Usage Map (FUM), is introduced to encode such dependencies so that the ranking of each product can be inferred using the features it possesses.

Another key challenge is that for a real CDSS, it is difficult to prepare a complete repository in advance for all possible intended usage. Maintaining an all-embracing knowledge base of FUMs for all using scenarios is neither economical nor feasible. In this paper, we adopt a query-driven strategy that allows the user to describe his or her intended usage with a short query keyword (or usage term), thereby permitting the system to derive the associated FUM in on-demand fashion. Such a “lazy” strategy defers case base induction and the model building process until the request for information is received. In contrast to conventional “eager learning” methods that require a complete training set containing all possible usage terms, this method avoids the headaches involved in maintaining a catch-all usage repository and training the FUMs in advance. In addition, it allows the existing usage term vocabulary to be incrementally enriched with system operation.

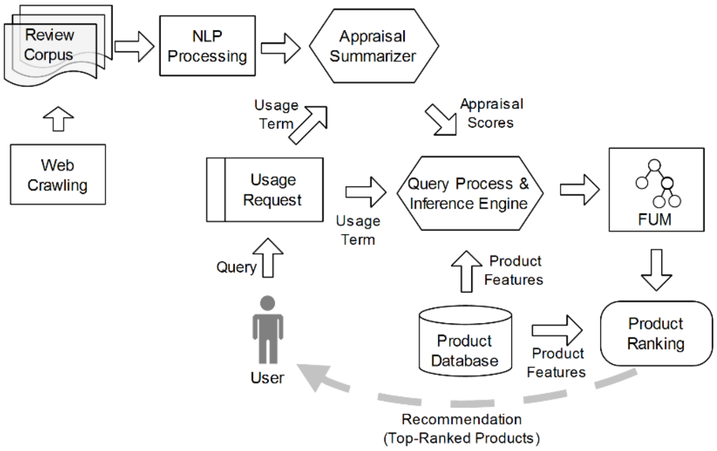


Figure 1: Architecture of the proposed framework—a query-driven strategy

The proposed product selection framework depicted in Figure 1 comprises the following major steps. 1) The system collects product review information by crawling the Web, and performs necessary preprocessing. 2) An online shopper specifies a usage term to start a query. The Appraisal Summarizer looks up query keyword from the annotated corpus and attempts to derive appraisal scores. 3) The Query Process and Inference Engine generates Training/testing cases by incorporating the product features stored in the product database and appraisal scores, and train an FUM regarding the given usage term. It also assesses the effectiveness of the obtained FUM (in terms of precision, recall, receiver operating characteristics (ROC), and the like). 4) If the FUM is effective, then it is applied to all product models in the product database to derive their appraisal ranking. Products with the top ranks are considered suitable for the specified usage, and thus suitable for recommendation to the user.

The digital camera purchasing scenario is used throughout this paper for illustrative purposes. However, the proposed framework, and its associated procedure and algorithms, are generalizable to many other application areas in which novice users need to choose products, services, or information in which their features play important roles in the decision.

The remainder of the paper is structured as follows. Section 2 provides a brief review of the relevant literature. FUMs, the key component of the framework, are formulated in Section 3. In Section 4 we introduce the Appraisal Summarizer that elicits product suitability class from product reviews. The procedures to handle query-based inference are detailed in Section 5. Section 7 introduces the prototype system based on the framework and exhibits its evaluation, and Section 8 concludes the paper with a summary and directions for future research.

2 LITERATURE REVIEW

Customer Decision Support. Popular recommendation techniques can induce appreciate products for customers, hence are widely applied to help make purchase decision. They mainly fall into two categories: the content-based approach and the collaborative approach. The content-based approach capitalizes on a customer's past purchases or rating information to predict his or her future purchase intention, whereas the collaborative approach recommends a product to the target customer based on analysis of other like-minded customers. Both approaches work well in recommending taste-related consumer products such as books or movies (Ricci, Rokach, Shapira, & Kantor, 2011), but neither is applicable to the context considered here because they rely on the users to express their preferences, either explicitly or implicitly, on various products in advance.

Review Summarization. To understand the appraisers' opinion from review texts, researchers have applied a number of feature extraction techniques to automatically identify the keywords of features or opinions. A number of well-established NLP approaches are useful in this regard, for example, part-of-speech (POS) tagging tools (Charniak, 1997) that can be used to identify the POS of words (e.g., adjective or adverb) in a review text. Some researchers have considered both product features and subjective terms when comparing products. For example, Red Opal (Scaffidi, et al., 2007) is an opinion mining-based product selection system that explores online customer reviews to identify product features and then automatically score products according to those features, thereby resulting in the most suitable product being recommended by matching products and features with those specified by the customer. Opinion mining techniques are also used for the automatic differentiation of the sentiment orientation (recommended or not) expressed toward an item in the text (Turney, 2002), which is particularly useful in helping purchase decision making. Some prior research has been done to Making recommendations based on opinion mining has also been studied by prior research including (Ku & Chen, 2007).

Product Recommendation based on eWoM. Researchers recognize that reviews and/or discussions of products on online forums and e-commerce websites have become important sources of product information (Dellarocas, 2006). There is also evidence to show that eWoM implied in online reviews can have a significant effect on customers' purchase decisions (Senecal & Nantel, 2004) and that such opinions can be exploited by intelligent systems to provide better recommendations. Many e-commerce websites and product review discussion boards provide ranking scores for various products, normally on a 5-point Likert scale, alongside the review text. However, recommendations based solely on overall ranking are problematic, as users' personal needs may differ from that of reviewers (Popescu & Etzioni, 2005). To address this problem, data mining and machine learning techniques, coupled with NLP approaches, have been developed to extract product ranking and other valuable information from product review texts, and are referred to collectively as opinion mining (Pang & Lee, 2008). (Sun, Long, Zhu, & Huang, 2009), for example, propose an intelligent recommendation approach based on the scores discovered in online reviews.

3 FEATURE USAGE MAP

The core of the framework is FUM, a graphical model that can be used to derive the suitability of a product (in terms of a certain usage scenario) based on the product features. This section formulates FUM and its two major presentations considered in this research.

Let $\mathcal{P}$ be a set of products of the same type described by feature set $\mathrm{F} = \{\mathrm{F}_1, \mathrm{F}_2, \dots, \mathrm{F}_\mathrm{n}\}$, where each F_i, i=1…n is a random feature variable. For example, product type "digital camera" has the feature "water tightness," which takes the value "yes" or "no." Target product $p \in \mathcal{P}$ can thus be represented by a vector

$p = \langle f_1, f_2, \dots, f_n \rangle$, where f_i is the value of F_i. Let S be a class variable that takes a value from $\{+, -\}$, indicating a "suitable" or "unsuitable" class, respectively. S represents the suitability of target product $p \in \mathcal{P}$ with reference to a specific product using scenario.

As previously noted, the suitability of a product in a particular using scenario is closely related to the features it possesses. Accordingly, to derive the suitability of a target product, we need to elicit the relations between feature variables $\{F_1, F_2, \dots, F_n\}$ and suitability variables S. In this research, the set of these relations are referred to as an FUM. The Bayesian network (BN) is reported to have a strong ability to model the probabilistic "cause-effect" relation between variables (Friedman, Geiger, & Goldszmidt, 1997), and is thus adopted in this research to constitute the FUM.

Assume that U is the set of all usage types. An FUM for any $u \in U$ is defined as a graphical model, $\mathrm{FUM_u} = (\mathrm{G}, \mathrm{P})$. $\mathrm{G} = (\{\mathrm{S}, \mathrm{F_1}, \mathrm{F_2}, \dots, \mathrm{F_n}\}, \mathrm{E})$ is a directed acyclic graph comprising a set of vertices $\{\mathrm{S}, \mathrm{F_1}, \mathrm{F_2}, \dots, \mathrm{F_n}\}$, each of which has a finite set of mutually exclusive states. E is the set of dependency relationships among the variables.

An unrestricted BN can take all of the dependencies between the feature variables into account. However, using it in the context under this study is infeasible because it is computationally difficult to establish a previously unknown network. When there are limited training data, especially, the complexity of an unrestricted BN may lead to a high degree of variance and thus to poor probability estimates (Cheeseman & Stutz, 1996). This paper adopts the tree augmented na ïve Bayes approach (TAN) (Friedman, et al., 1997), which relaxes the independence assumption in the Na ïve Bayesian Network by allowing each independent variable to have at most one non-dependent parent.

Given an FUM defined on variable set $N = \{\mathrm{F_1}, \mathrm{F_2}, \dots, \mathrm{F_n}, \mathrm{S}\}$ with a determined structure, let $\pi(\mathrm{F}_j)$ denote the set of parents of $\mathrm{F}_j \in \mathrm{N}$. A TAN-assembled FUM thus needs to satisfy the following constraints. (1) $\pi(\mathrm{S}) = \emptyset$; (2) $\mathrm{S} \in \pi(\mathrm{F}_j)$ for all $1 \le \mathrm{j} \le \mathrm{n}$; and (3) $|\pi(\mathrm{F}_j)| \le 2$ for all $1 \le \mathrm{j} \le \mathrm{n}$. Constraint (1) requires the suitability variable/node to be the root; (2) means that every feature variable F_j must have parent node *S*; and (3) allows any feature variable F_j to have at most one non-root parent.

According to the Bayesian theorem, the conditional probability of product $p = \langle f_1, \dots, f_n \rangle$ with suitability s can be calculated by $\Pr(S = s|p) = \frac{\Pr(S=s)}{Pr(p)} \prod_{j=1}^{n} \Pr(f_j | \pi(f_j))$.

If we view an FUM as function mapping $FUM_u\colon F \to S$, because the state of S is irrelevant to $Pr(f_1, f_2, \dots, f_n)$, the most probable suitability class regarding usage u is $\mathrm{FUM_u(p)} = \underset{s \in \{+,-\}}{\operatorname{argmax}} \Pr(s) \prod_{j=1}^{n} \Pr(f_j | \pi(f_j))$, where $\Pr(\mathrm{s}) = \frac{N_i}{N_0}$, and where N_i is the number of products suitable for u_i. $N_0 = \sum_{i=1}^{|U|} N_i$ is the total number of products appearing in the dataset.

This research learns FUMs based on product features and available suitability information. Learning TAN comprises two main steps, that is, structure learning and parameter learning. The former step can be performed by finding the maximum weight-spanning tree using the Chow and Liu algorithm (Chow & Liu, 1968), whereas the latter involves calculation of the joint probability distributions. More specifically, in the FUM defined above, if f_j has only one parent u_i, we can calculate $\Pr(f_j | \pi(f_j)) = \frac{N_{ij}}{N_j}$, where N_{ij} is the number of products suitable for u_i and has feature state f_j. Otherwise, that is, if f_j has two parents u_i and $f_{k,}$, then $\Pr(f_j | \pi(f_j)) = \frac{N_{ijk}}{N_{ik}}$, where N_{ijk} is the number of products suitable for u_i and has feature states f_j and f_k, and N_{ik} is the number of products suitable for u_i and has feature state f_k. To tackle the zero-probability problem, a smoothing method needs be used to adjust the foregoing calculation.

4 APPRAISAL SUMMARIZER

The appraisal summarizer derives the suitability class of products regarding a specified product usage from review corpus. This process consists two main steps. (1) We build an Appraisal Classifier (AC) to identify appraisal sentences, which are subjective sentences that comment on a given product usage, and (2) we apply sentimental analysis techniques to the appraisal sentence to derive the suitability label of the product. Due to the page limitation, the details of the first step are omitted in this paper. In the second step, the suitability label are derived through sentiment information, which is the polarity of the reviewers with respect to using the target product in a particular scenario. This task is accomplished with the following procedure.

(1) Calculating the polarity of lemmas. The major linguistic resource adopted is SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010), a lexical resource that is widely used to process natural language to better understand sentiment terms. Polarity information in SentiWordNet is quantified on the basis of the lexica in WordNet using linguistic and statistical classifiers. A synset in SentiWordNet is associated with three polarity scores (positivity, negativity, and objectivity), and the sum of the three equals 1. For instance, the triplet (0, 0.75, and 0.25) (positivity, negativity, and objectivity) is assigned to the lemma “poor”. Note that a lemma in SentiWordNet may belong to multiple synsets that may have different positive/negative polarity scores. As (Neviarouskaya, Prendinger, & Ishizuka, 2009) suggest, the positive (negative) polarity of a lemma can be calculated by averaging all non-zero positive (negative) scores for its corresponding POS.

(2) Deriving the polarity of an appraisal sentence. When calculating the overall appraisal score of a sentence, we consider its S-V-O-P segment (subject phrase, verb phrase, object phrase, and prepositional phrases), which is a subtree in the syntactic parse tree, rather than the whole sentence to more accurately focus the calculation on the target product. For example, in the sentence “In contrast to the laggy autofocus of its predecessor, G12’s improved autofocus does a very good job of nailing focus quickly in low light condition,” only the fragment [G12’s improved autofocus][does][a very good job of nailing focus quickly] is counted. We use a modified version of the term counting method (Kennedy & Inkpen, 2006) to form a composite of the overall polarity in which the average polarity score of all adjectives and adverbs appearing in the fragment is calculated. This method requires no training, yet has been reported to achieve a level of accuracy close to that of the supervised learning approach (Kennedy & Inkpen, 2006). A further enhancement of this method is to take intensifiers (i.e., “very”) and diminishers (i.e., “barely”) (Kennedy & Inkpen, 2006) and the scores of adverbial modifiers into account. During the calculation, a negation term (such as “not,” “never,” or “none”) reverses the corresponding score of an adjective/adverb. For instance, the phrase “not good” has a polarity score of -0.75 if the polarity score of “good” is “0.75.” The overall polarity score of an appraisal sentence s in which a specified usage term u occurs is denoted as $polarity(u, s_u)$, which takes a value from the interval [-1, +1].

(3) Summarizing suitability score. The suitability score of a product can thus be calculated as the average polarity of all appraisal sentences (with regard to the given usage) in the review documents about product p. that is, $score(u,p) = \frac{1}{\sum_{d_i \in D_p} |d_i|} \sum_{d_i \in D_p} \sum_{s_u \in d_i} polarity(u, s_u)$, where s_u is an appraisal sentence containing term u, and $|d_i|$ is the number of such appraisal sentences occurring in document d_i. D_p is the set of all reviews about product p. Hence, the case with sufficiently high suitability score should be considered as “suitable”. That is, the class label of a case associated with product p and term u is determined by a predefined cut-off, or $class(u,p) = \begin{cases} + & score(u,p) \geq \tau \\ - & otherwise \end{cases}$.

5 DERIVING RECOMMENDATIONS

A case base is the collection of training/testing instances used in building FUMs. A product model is corresponding to a single case, which comprises the feature attributes pertaining to the product and a class attribute (the suitability of the product in terms of a specified usage), denoted $c = \langle f_1, f_2, \ldots, f_n, s \rangle$. This section elaborates the procedure to process a user query. The procedure completes three tasks, i.e, generating the case base, training the FUM, and applying the FUM to derive recommendations, which are illustrated in the following two algorithms.

5.1 Case Base Generation

Given review corpus *D*, usage term *u*, and product database *PD* as the inputs, the algorithm in Figure 2 outlines the case base generation procedure. The algorithm locates all products that have been reviewed (regarding the given usage term) before, and each of which corresponds to a case (lines 1). For each product, its feature attribute can be retrieved from product database *PD* (line 3), and the class attribute can be derived using the aforementioned appraisal summarizer. Hence, a new case can be generated by merging the feature attributes and class attributes. The new case obtained can then be added to *CB*u, the case base for usage term u.

```
   Input: review corpus D, usage term u, product database PD
   Output: case base associated with u:CBu
   Method:
1  $PD_u = \{p \in PD | \ p$ has review document in $D$ regarding usage term $u$ $\}$
2  for each $p \in PD_u$ do
3      $\langle f_1, f_2, \dots, f_n \rangle \leftarrow Retrieve\_Feature(PD, d)$
4      $s \leftarrow class(u, p)$
5      $c_d \leftarrow \langle f_1, f_2, \dots, f_n, s \rangle$
6      $CB_u \leftarrow CB_u \cup \{c_d\}$
7  Return $CB_u$
```

Figure 2: Algorithm 1—case base construction

5.2 Training FUM and Deriving Recommendations

In fact, for all product in case base CB_u, their summarized suitability scores regarding u are readily available. Hence a straightforward recommendation method is to advise those $c \in CB_u$ with top summarized suitability scores. However, this case base only involves those product models been reviewed before, recommendation based on such an incomplete product information source will therefore be biased. In contrast, the method introduced in this subsection accounts all models in product database when inducing recommendations.

The on-demand strategy used in case base construction allows the FUM base to be built in an incremental fashion. Suppose that the FUMs corresponding to usage terms $U = \{u_1, u_2, \dots, u_i\}$ have previously been obtained, and are denoted $UM_U = \{FUM_{u1}, FUM_{u2}, \dots, FUM_{ui}\}$. The following algorithm (Figure 3) describes the procedure for processing a query with usage term *u*. It attempts to derive an FUM corresponding to the given usage term and apply it to spot the highest ranked product models in the database.

```
   Input: $FUM_U = \{FUM_{u1}, FUM_{u2}, \dots, FUM_{ui}\}$, product feature database PD, review corpus D,
   Output: top-j ranked products regarding usage u
   Method:
1  if $u \in U$ , then return $top\_j(FUM_u, PD, u)$;
2  else
3          $CB_u \leftarrow construct(PD, D, u)$
4          k-fold training and testing to induce $FUM_u$
5          if is_good$(CB_u, FUM_u)$, then
6                  $FUM \leftarrow FUM \cup \{FUM_u\}$; $U \leftarrow U \cup \{u\}$
7                  return $top\_j(FUM_u, PD, u)$;
8          else  return NULL
```

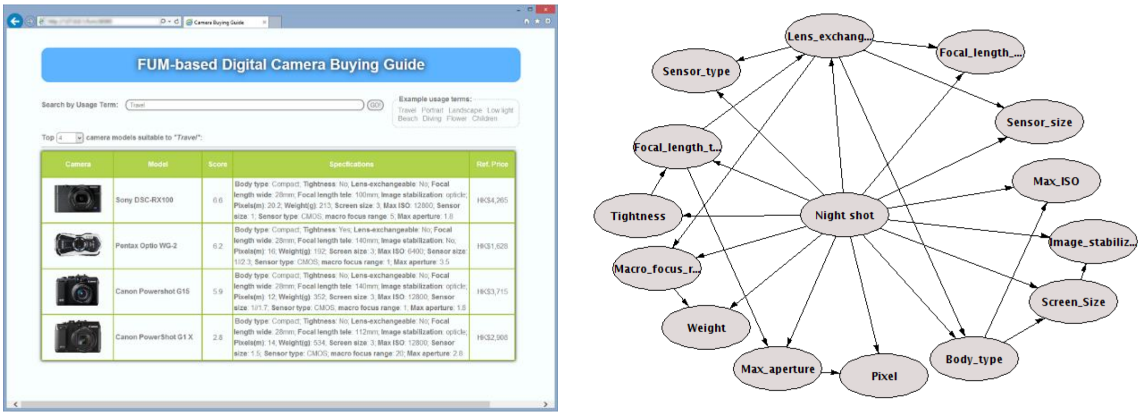
Figure 3: Algorithm 2—query-driven training, testing of FUMs and products scoring

When the user starts a query with term *u*, the algorithm looks it up in *U*, which is the set of usage terms whose FUMs have been trained and are ready to use. If *u* is found, the algorithm directly applies its corresponding map FUM_u to the product database to locate the top-j highly ranked products in terms of *u* (line 1). Note that instead of retrieving from a stored top product list, this extraction should be invoked every time to allow the most up-to-date products to be taken into consideration. If *u* is a previously unseen usage term, this algorithm invokes algorithm 1 to construct cases in accordance with *u*. Consequently, it trains FUM_u using the obtained cases and assesses its overall performance. The new classifier FUM_u can be stored if it demonstrates satisfactory performance (in terms of such criteria as precision, recall, F-value, and area under the ROC curve [AUC]), and hence *u* is marked as “trained” (line 6). As such, the FUM base grows each time when the system handles a new query request. This new classifier associated with the input usage term *u* can thus be applied to the product feature base to induce the top-*j* highest ranked products in terms of their posterior odds of the products regarding the given FUM$_u$ defined by $\mathrm{SO}(\mathrm{p},\mathrm{u}) = \frac{\Pr(S=+|p)}{\Pr(S=-|p)} = \frac{\Pr(S=+)}{\Pr(S=-)} \prod_{j=1}^{n} \left(\frac{\Pr(f_j|\pi^+(f_j))}{\Pr(f_j|\pi^-(f_j))} \right)$, where p is the target product and u is the usage. Otherwise, if the induced FUM is found not with sufficiently classification performance, a “NULL” recommendation will be given (line 8).

6 EVALUATION

As a manifestation of the motivating scenario, we choose digital camera as the example product category for illustration. A prototype system was developed in order to demonstrate the feasibility of applying the proposed framework and associated algorithms in building a real Customer Decision Support System for assisting digital camera purchase. We developed the Appraisal Summarizer based on GATE (Cunningham, 2002), an open source platform providing general text processing workflows for solving common NLP problems. The Coreference resolution was also implemented using the “Orthomatcher tool” provided in GATE. The product database and review corpus are the core of the system’s operation. The former was created based on product data collected from the website Dpreview.com and the latter was built using over 220,000 customer reviews scraped from Amazon.com. The basic query-driven processing logic was developed using Weka (Hall, et al., 2009).

A screenshot of the user interface of the prototype system is shown in Figure 4(a), while Figure 4(b) presents an example FUM for the using scenario “night shot” trained using real data from the product database.



FUM-based Digital Camera Buying Guide

Search by Usage Term: Travel GO!

Example usage terms: Travel Portrait Landscape Low light Beach Diving Flower Children

Top 4 camera models suitable to *"Travel"*:

Camera	Model	Score	Specifications	Ref. Price
	Sony DSC-RX100	6.6	**Body type**: Compact; **Tightness**: No; **Lens-exchangeable**: No; **Focal length wide**: 28mm; **Focal length tele**: 100mm; **Image stabilization**: opticle; **Pixels(m)**: 20.2; **Weight(g)**: 213; **Screen size**: 3; **Max ISO**: 12800; **Sensor size**: 1; **Sensor type**: CMOS; **macro focus range**: 5; **Max aperture**: 1.8	HK$4,265
	Pentax Optio WG-2	6.2	**Body type**: Compact; **Tightness**: Yes; **Lens-exchangeable**: No; **Focal length wide**: 28mm; **Focal length tele**: 140mm; **Image stabilization**: No; **Pixels(m)**: 16; **Weight(g)**: 192; **Screen size**: 3; **Max ISO**: 6400; **Sensor size**: 1/2.3; **Sensor type**: CMOS; **macro focus range**: 1; **Max aperture**: 3.5	HK$1,628
	Canon Powershot G15	5.9	**Body type**: Compact; **Tightness**: No; **Lens-exchangeable**: No; **Focal length wide**: 28mm; **Focal length tele**: 140mm; **Image stabilization**: opticle; **Pixels(m)**: 12; **Weight(g)**: 352; **Screen size**: 3; **Max ISO**: 12800; **Sensor size**: 1/1.7; **Sensor type**: CMOS; **macro focus range**: 1; **Max aperture**: 1.8	HK$3,715
	Canon PowerShot G1 X	2.8	**Body type**: Compact; **Tightness**: No; **Lens-exchangeable**: No; **Focal length wide**: 28mm; **Focal length tele**: 112mm; **Image stabilization**: opticle; **Pixels(m)**: 14; **Weight(g)**: 534; **Screen size**: 3; **Max ISO**: 12800; **Sensor size**: 1.5; **Sensor type**: CMOS; **macro focus range**: 20; **Max aperture**: 2.8	HK$2,908

Figure 4: The sub-figure on the left exhibits a screenshot of the Web-based user interface of the prototype system; the sub-figure on the right is an example FUM with the usage “night shot”.

An experiment was conducted in order to examine the usefulness of the system based on the proposed framework in a real online shopping scenario. We experimented with the following approaches in a simulated product selection scenario: (1) Non-prosumers actually study product specifications and reviews to identify a most suitable product model, referred to as “Manual Approach” hereafter. This approach was used as the baseline for comparison. (2) Summarizing suitability scores and recommending the products with top scores (mentioned in section 5.2), referred to as “SUM Approach”, and (3) The FUM-based approach described in algorithm 2, referred to as “FUM Approach”.

Specifically, the following two hypotheses that are relevant to this research were examined. We believe that the products selected by the proposed FUM-based approach will be more suitable (in terms of the given usage term) than those selected by actual non-prosumers. First, it is, in general, difficult for non-prosumers to understand product specifications and jargons mentioned in reviews, hence the purchase decision based on them could be misdirected. Second, it has be recognized that the set of alternative products that a purchaser may take into consideration is limited, because the expected utility for further information search decreases as more products are examined (Stigler, 1961). In contrast, the proposed FUM approach can exhaust all available product information and their comments and based on which to derive recommendation, hence is expected to be able to make better purchase decision. Therefore, we have the following hypothesis:

Hypothesis 1 (H1): Products recommended by FUM-based approach has higher suitability than those by manual selection.

In addition, we argue that the FUM approach can produce better recommendations than SUM approach in general, since SUM approach can only induce appraisal scores of products been reviewed before. As a step further, the FUM approach is capable of learning the Feature-Usage Map from existing cases and applying it to the whole population of the product database to infer the optimal choices. Hence the following hypothesis:

Hypothesis 2 (H2): Products recommended by FUM-based approach has higher suitability than those by SUM-based approach.

The user study was on a web-based platform that we developed specifically for conducting the experiment. This platform can also invoke the prototype system automatically to induce recommendations for making comparison. The empirical evaluation carried on in this study was based on domain experts' judgment for assessing the quality of purchasing decision. Due to space constraints, the simulation to evaluate the technical performance is not discussed in this paper.

6.1 Participants and Procedures

An advertisement was posted on a popular online discussion forum to invite voluntary participants. Totally 129 respondents took part in our web-based user study. Phases 1 is used to screen out the prosumer participants, because the design of the system was targeting on non-prosumers only. Valid non-prosumer participants were requested to further complete the product selection task in phase 2. In phase 3 the platform invoke the prototype system to generate recommendations for comparison. Finally, in phase 4, domain experts evaluated and rated the quality of the selected products.

Phase 1 (pre-task questionnaire and screening): After signing a consent form, participants were requested to answer several extra questions used to measure the level of expertise of the participants so that prosumer participants can be screened out. The questionnaire and criteria for screening were developed by domain experts: 5 questions related to digital photography equipment selected from "Certified Photographer Exam Papers"[3] were added into the pre-task questionnaire in order to measure the "richness" of a participant's knowledge about digit photography and thus identify those prosumer participants. In the end, participants who answered 3 or more questions correctly were considered prosumers thus would not continue to complete the questionnaire, while the rest were requested to fill out the general questionnaire about demographic. As a result, 52 participants remained after the screening.

Phase 2 (user task): An information-seeking task was given to the participants so as to measure their performance in studying web resources and making purchase decision. Prior research (Borlund, 2000) has affirmed that an exploratory information-seeking task in experiment can reflect a user's real-life information needs if it is phrased as a situational task in a simulated scenario. Since the market of digital camera is also vertically differentiated, two cameras are comparable, especially in terms of performance in building quality, durability, and the like, only when they are on the same price level. Based on domain experts' suggestion, therefore, we considered three price ranges (i.e., $"\leq 3{,}000"$, $"\text{between } (3{,}000, 4{,}500]"$, and $"\geq 4{,}500"$, in Hong Kong Dollar) that correspond to three most typical tiers (entry-level, mid-range, and high-end, respectively) of digit camera. The task had a two-step procedure:

Step 1: For each participant, the experiment platform first randomly chose a range from the 3 price ranges. The participant was then given the task description, in which s/he was awarded a bounty amounting to the upper bound of the chosen range, and requested to purchase a new camera priced in the range with it. The participants were asked first to envisage the scenario in which they will use the new camera, and then describe it with a short phrase no more than two words like "travel", "low light", etc. and input it into the search box on the webpage for experiment.

[3] http://www.certifiedphotographer.net, retrieved in May 2013.

Step 2: Subsequently, the participants were required to explore reviews and specifications on two imposed websites Dpreview and Amazon for spotting the most suitable camera model to their using scenario with the given budget. The specifications-based product purchasing guiding tools in these two websites were recommended to the participant to facilitate task completion. The participants were instructed to report the selected camera once s/he made the purchase decision. In particular, if the participant found unable to complete the task, s/he was allowed to terminate without reporting his/her selection. In such a case, his/her selection would be marked “NULL” by the system automatically.

Phase 3 (Post-task automatic process): In this phase, our experiment platform invoked the SUM and FUM procedures to select the most suitable models. For each participant, its corresponding budget range and usage terms were passed to both SUM and FUM approaches. Subsequently, the SUM approach derived the top-ranked camera whose prices were within the designated budget using its summarized suitability score in terms of the given usage term. Likewise, FUM identified the top-ranked camera using posterior odds score (SO). As a result, each manual selection from a participant corresponded to two algorithm-generated selections by SUM and FUM approaches, respectively. Specifically, in the case that an approach was unable to make a selection with the given usage term and budget, its result was labelled “NULL”.

Phase 4 (Expert Evaluation): After all replies were collected and all machine-based recommendations were generated, three domain experts were invited to manually assess the suitability level of camera models recommended by three different approaches, i.e., “Manual”, “SUM”, and “FUM”, for the given usage and budget range. For each participant with specific usage term and budget, the selected product models by three approaches were presented side by side to facilitate the experts to review and score. A five-point Likert scale ranging from 1 (“least suitable”) to 5 (“most suitable”) was used to evaluate the suitability level. In particular, we considered the results labelled “NULL” the least suitable, hence would be scored 1 by the experiment system automatically. The overall evaluation score of a specific recommendation was calculated by averaging the scores from three different experts.

6.2 Results and Discussion

The summary statistics for the experiments results are presented in table 1. We conducted repeated measures ANOVA and found that, the difference in evaluation scores between the Manual Approach, the SUM Approach, and the FUM Approach was statistically significant, F(2, 51) = 8.007, p < .01. The pairwise comparison further showed that, the suitability evaluation for the FUM Approach was significantly higher than that for the Manual Approach (3.115 vs. 2.467, p < .001), which supported H1. Meanwhile, the difference in scores between the FUM Approach and the SUM Approach was also significant (3.115 vs. 2.717, p < .05), and H2 was supported. Therefore, the FUM approach, evidently, outperformed the other two approaches in terms of the suitability of the recommendations.

The average suitability score of 3.115 (of 5) for the recommendations yielded by FUM does not appear to be an outstanding result in a general sense. This score, however, is rather a relative criterion than an absolute performance indicator of recommendation effectiveness. Because raters tend to be in favor of their most familiar products (the Mere-exposure effect) while the recommendations involve a large number of camera models. Therefore, such a bias leads to severity tendency on the rating scores.

Of all the 52 recommendations produced by the participants (Manual Approach), 12 (i.e., 23.08%) were “NULL”, which means 12 (non-prosumer) participants were unable to complete the given task. The number of “NULL”s for “SUM Approach” and “FUM Approach” was 1 and 6, respectively, accounting for 1.92% and 11.54% of all recommendations. Owing to the large size of the used review corpus, notably, the SUM can almost always produce a recommendation for any given usage terms, albeit the average suitability of its recommendations was not the highest.

Approach	# of NULL	% of NULL	Mean	SD
Manual	12	23.08%	2.467	0.966
SUM	1	1.92%	2.717	0.909
FUM	6	11.54%	3.115	0.973

Table 1. Summary statistics for the experiments results. The second column shows the total number of “Null recommendation” produced by the corresponding approach. The third

column exhibits the percentage of "Null" (over a total of 52 recommendations) produced by the corresponding approach.

Although the recommendation quality of the SUM approach was much lower than that of FUM, it produced much less null recommendations, because the FUM approach has no restriction on sample size. Hence, an immediate enhancement to reduce the number of null recommendations is to exploit the ensemble of the FUM and SUM approaches. In other words, when the system fails to yield a qualified FUM, it retrieves the existing case base (CB_u in algorithm2) and recommends those products with top suitability scores instead.

Notably, we have observed that FUMs derived by popular usage terms (such as "beginner") generally perform better than by rare ones do (e.g., "amateur"). It is because an infrequent term often results in smaller training sample size that may lead to a classifier with higher error rate. A straightforward improvement in a real-world CDSS is to allow the user to choose the usage term among the original term's synonyms, hypernyms, and hyponyms in an interactive manner, or combine their generated training sets, until a sufficiently accurate FUM is obtained.

7 SUMMARY AND FUTURE RESEARCH

This paper presents a generic framework with detailed procedures for assisting non-prosumers to make purchase decision with minimal information requirement, i.e., a short term describing the using scenario. The framework represents the relationships between a product's suitability in terms of a specific using scenario and its technical specifications (features) using a Bayesian Network referred to as FUM. A product's suitability score can thus be derived from the features it possesses. The paper also proposes an appraisal summarizer that elicits implicit product rating from eWoM (i.e., product reviews) for automating case base generation for the training and validation of FUMs. In addition, a generic query-based procedure that builds case bases, trains and validates the FUMs, and applies them to infer product recommendations based on the given usage term is elaborated. To evaluate our framework and the associated algorithms, two types of evaluations were conducted. The simulation-based evaluation confirmed that a TAN-based FUM generally outperforms a NB-based FUM when used to classify products according to their suitability. This results suggest that when used to classify differentiated products based on their features, a TAN-based FUM achieve fine classification performance while remain moderate model complexity, hence is a fair option for CDSS. Furthermore, an empirical evaluation was also conducted in order to examine the effectiveness of the prototype CDSS when used by non-prosumers in a simulated purchase decision-making scenario. Its results show that the prototype system can help novice consumers make better purchasing decisions than conventional feature-based systems do. The experimental results not only demonstrate the practical relevance of the proposed framework and associated algorithms, but also lead to several potential alterations that can fit the system into a real world application. Other than the customer decision-support domain, the proposed framework, alongside its associated procedures and algorithms, may have extensive applications in other areas where online reputation of the target items to be recommended plays a central role.

In our future research, we will focus on developing a full-fledged CDSS by enhancing the framework in the following two aspects. First, it has been observed that the low-quality reviews are common on the Web (Leea & Choeh, 2014), which may potentially jeopardize the performance of the CDSS or even cause it fail in practice. In our current implementation of the prototype system, the problem is alleviated by ruling out low-rating reviews from the Amazon review corpus. Nevertheless, a more generalizable and robust method to deal with low-quality reviews is crucial, since ratings of reviews are not readily available in general. A viable solution is to develop a document-level classifier according to reviews' quality. Some recent studies have already shed light on this area (Dey & Haque, 2009; Leea & Choeh, 2014). Additionally, another relevant practical extension of the framework is to allow multiple using scenarios. A straightforward approach is to incorporate the posterior odds into the Weighted Sum Model (Fishburn, 1967). That is, the consumer is requested to specify a weight for each given using scenario. Hence, a separated FUM will be built for each using scenario, thus a utility score of a product that represents the product's overall suitability rank can be considered as the weighted sum of the Suitability Odds derived by each individual FUMs.

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Language Resources and Evaluation*. Valletta, Malta.

Borlund, P. (2000). Experimental components for the evaluation of interactive information retrieval systems. *Journal of documentation, 56*, 71-90.

Charniak, E. (1997). Statistical techniques for natural language parsing. *Ai Magazine, 18,* 33.

Cheeseman, P., & Stutz, J. (1996). Bayesian Classification (AutoClass): Theory and Results. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*: AAAI Press/MIT Press.

Chen, S.-C. (2011). Understanding the effects of technology readiness, satisfaction and electronic word-of-mouth on loyalty in 3C products. *Australian Journal of Business and Management Research, 1,* 1-9.

Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on, 14*, 462-467.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities, 36*, 223-254.

Dellarocas, C. (2006). Strategic manipulation of Internet opinion forums: Implications for consumers and firms. *Management Science, 52*, 1577-1593.

Dey, L., & Haque, S. K. M. (2009). Opinion Mining from Noisy Text Data. *International Journal on Document Analysis and Recognition, 12*, 205-226.

Fishburn, P. C. (1967). *Additive Utilities with Incomplete Product Set: Applications to Priorities and Assignments*. Baltimore, MD, U.S.A.: Operations Research Society of America (ORSA).

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning, 29,* 131-163.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*.

Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence, 22*, 110-125.

Kotler, P., & Keller, K. L. (2006). *Marketing Management* (12 ed.). New Jersey: Prentice Hall.

Ku, L. W., & Chen, H. H. (2007). Mining opinions from the Web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology, 58*, 1838-1850.

Leea, S., & Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications, 41,* 3041–3046.

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2009). Sentiful: Generating a reliable lexicon for sentiment analysis. In *Affective Computing and Intelligent Interaction and Workshops, 3rd International Conference on* (pp. 1-6): IEEE.

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval, 2*, 1-135.

Popescu, A. M., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 339-346). Stroudsburg, PA.

Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). *Recommender systems handbook*: Springer.

Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H., & Jin, C. (2007). Red Opal: product-feature scoring from reviews. In *the 8th ACM conference on Electronic commerce* (pp. 182-191). San Diego: ACM.

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing, 80*, 159-169.

Stigler, G. J. (1961). The Economics of Information. *The journal of political economy, 69*, 213-225.

Sun, J., Long, C., Zhu, X., & Huang, M. (2009). Mining Reviews for Product Comparison and Recommendation. *Research journal on Computer science and computer engineering with applications,* 33-40.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *the Meeting of the Association for Computational Linguistics (ACL)*.