On Argument Bundles in the Web of Experiences

Xavier Ferrer[1,2] and Enric Plaza[1]
[1] *Artificial Intelligence Research Institute (IIIA), Spanish National Research Council (CSIC), Campus U.A.B., Bellaterra, Catalonia (Spain)*
[2] *Universitat Autònoma de Barcelona, Bellaterra, Catalonia (Spain)*
{xferrer,enric}@iiia.csic.es

In this paper we focus on a particular interesting type of web user-generated content: people's experiences. We extend our previous work on aspect extraction and sentiment analysis and propose a novel approach to creating a vocabulary of basic level concepts with the appropriate granularity to characterize a set of products. This concept vocabulary is created by analyzing the usage of the aspects over a set of reviews, and allows us to find those features with a clear positive and negative polarity to create the bundles of arguments. The argument bundles allow us to define a concept-wise satisfaction degree of a user query over a set of bundles using the notion of fuzzy implication, allowing the reuse of experiences of other people to the needs a specific user.

Keywords: Web of Experiences, sentiment analysis, arguments, aspect extraction, basic level concepts

1. Introduction

Our work is developed in the framework of the Web of Experiences [10]. This framework proposed to enlarge the paradigm of Case-based Reasoning (CBR), based on solving new problems by learning from past experiences, and includes all forms of experiences about the real world expressed as user-contributed content on the web. The final goal is to *reuse* this collective experience in helping new people (the "users") in taking a more informed decision according to their preferences, which can be different from the preferences of the individuals who have expressed their experiences on the web. The overall goal of the Web of Experiences (WoE) approach is constructing the relationship between these two points: from numerous but diverse *individual experiences* to a *specific and personal user request*, and this paper presents a complete instantiation of this relationship.

In this WoE approach we focus on praxis and usage, and we want to analyze how individuals express certain experiences about their daily life; in this paper we will focus on the usage of digital cameras. A main goal is to discover the vocabulary they use, which do not need to be the same as the classical feature list describing the different aspects of a camera (e.g. "has 45 AF points"), in order to elucidate the main pros and cons of each camera according to the user reviews.

To this end, we analyze textual reviews of user experiences with digital cameras and identify the set of aspects the individuals utilize and the polarity of the sentiment words associated with them [15,16]. Aspects are grouped in basic level concepts (BLC) [11], creating a new *concept vocabulary*, to overcome the disparate granularity of the extracted aspects. Those concepts with a strong positive or negative polarity over the set of reviews of a camera are considered the *pros and cons*, respectively, of that camera.

We call a *bundle of arguments* the set of main pros and cons of a camera. We take this approach, already envisioned in [10], because the pros and cons allow us to acquire and reuse the knowledge for other people with diverse individual preferences. To support this reuse, we introduce the notion of query satisfaction by a bundle of arguments modeled using the notion of fuzzy implication, in which a query expresses the new user individual preferences (e.g. a travel photographer will strongly prefer a camera with long battery life).

The paper is organized as follows: Section 2 describes aspect extraction and the discovery of basic level concepts from user reviews. Section 3 presents three different types of argument bundles and and Section 4 defines the user query satisfaction degree. Evaluation results are presented in Section 5, related research in Section 6, and conclusions in Section 7.

2. Aspects and Basic Level Concepts

In our previous work on social recommender systems we harnessed knowledge from product reviews, and characterized every product by a set of aspect-sentiment pairs extracted from its reviews [15]. Based on these characterizations, we ranked and selected the most useful aspects for recommendation [16]. However, even after identifying the most useful aspects for recommendation, we still processed synonymous aspects and aspects referencing the same concept (such as *sensor* and *cmos*) as different aspects, adding noise to the recommendation process.

In this work, we use a similar approach to [15] in order to extract the set of salient aspects used to define important characteristics of photographic digital cameras. We call *aspect vocabulary* $\mathcal{A}$ the set of extracted aspects from the reviews of a corpus $\mathcal{K}$, and $\mathcal{P}$ the set of products described in $\mathcal{K}$ reviews. However, instead of characterizing the products of the corpus directly by the aspect vocabulary, we group them in *basic level concepts* (BLC). According to Rosch et al. [11], basic level concepts are those that strike a tradeoff between two conflicting principles of conceptualization: inclusiveness and discrimination. Rosch et al. found that there is a level of inclusiveness that is optimal for human beings in terms of providing optimum cognitive economy. This level of inclusiveness is called the basic level, and concepts or categories at this level are called basic-level concepts. In Figure 1, the basic level is found at the middle level of the hierarchy, and the basic level concepts are 'picture', 'battery' and 'lens'. The concept 'image' is the superordinate concept for 'picture' (as are 'energy storage' and 'optics' to their respective BLCs), and it is more general: a 'picture' is considered here an *image made using a camera*, and used in common parlance more often than the technical term 'photograph'; moreover, 'image' is considered an abstract term, a *physical likeness or representation of a person, animal, or thing*, including painted canvases, and sculptures. 'RAW picture' is a subordinate concept of the BLC 'picture', and therefore it is more specific. Notice that the BLC subordinate concepts are usually formed by combining the BLC concept name with another word that specifies the nature of the given BLC. For instance, 'RAW picture', or other concepts such as 'gray-scale picture' and 'color picture' are also subordinate concepts of the BLC 'picture', because they are more specific and can be abstracted to have the same intended meaning as 'picture'.

Research in the field of identifying basic level concepts is mostly oriented to improve the task of *word sense disambiguation*. For instance, the class-based word sense disambiguation [7] approach requires to manually identify words in a corpus as belonging to

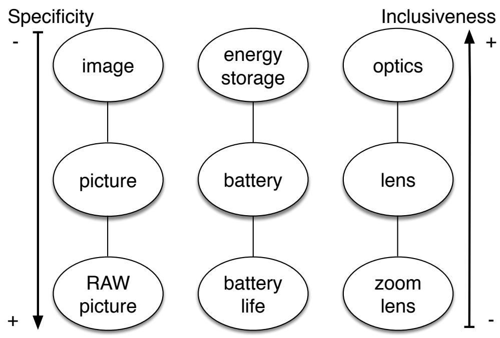

Fig. 1. Examples of several levels of categorization on digital cameras.

one semantic class, that is interpreted as one BLC. Once the corpus is marked, several supervised classifiers are trained to assign the proper semantic class to each ambiguous word. In our approach, we discover a collection of basic level concepts in an unsupervised way from the review corpus, where each BLC assembles a set of aspects that, according to our analysis, are used in a similar way by the reviewers. As we show in section 2.1, we estimate this similarity by taking into account semantic similarity and evaluating the coherence/incoherence of the sentiment values of the aspects assembled in a given BLC. Synonymy is a special case of aspects *being used in a way that is indistinguishable* for our purposes.

Consider, for instance, these three aspects in $\mathcal{A}$: *picture*, *pic* and *jpeg*. One may surmise people using those words in reviews are in fact referring to the same basic level concept, i.e. *the picture obtained by my digital camera*. Thus, we could consider that different reviews in the corpus using those words are referring to the same BLC, because they have the same intended meaning (they are indistinguishable for our purposes).

We will now present a method to discover a collection of BLCs in order to to create a concept vocabulary $\mathcal{C}$. The creation of a collection of basic level concepts consist of three steps: 1) identifying synonymous aspects, 2) building a hierarchical clustering using a new similarity measure among aspects, and 3) creating a concept vocabulary $\mathcal{C}$ of basic level concepts from the hierarchical clustering.

2.1. Hierarchical Clustering of Aspects

The first step is, considering a corpus $\mathcal{K}$ with an aspect vocabulary $\mathcal{A}$, to identify which aspects in $\mathcal{A}$ are synonyms according to WordNet (a lexical database of English). To do so, we first need to identify the Word-

Net synsets of every aspect and identify if they are related. The five more frequent aspects of $\mathcal{A}$ are manually mapped to the corresponding WordNet synset related to digital photography. For each of the remaining non-mapped aspects of $\mathcal{A}$, we first search the corresponding WordNet *synsets* (if any) associated with the noun word form, and then disambiguate it by identifying the *synset* with the shortest aggregated WordNet *Path Distance* [8] with respect to the five previously manually selected synsets. The aspects that have a synonymy relation among them are assembled into *aspect groups* $\{G_1,\ldots,G_m\}$. Aspects with no synonyms form a singleton group.

Next, we iteratively cluster the most similar groups of aspects and create a dendrogram. To cluster the aspect groups we use an unsupervised bottom-up hierarchical clustering algorithm that takes the most similar pair of groups at each stage and puts them together in a higher level group.

We will now define similarity measures over aspects and over groups. The similarity measure between two aspects a_i and a_j is:

$$Sim_A(a_i,a_j) = \alpha\cdot\Gamma(a_i,a_j)+\beta\cdot\Phi(a_i,a_j)+\gamma\cdot\Lambda(a_i,a_j)$$

where α, β and γ are weighting parameters in $[0,1]$ such that $\alpha+\beta+\gamma=1$. The values of Sim_A are in $[0,1]$. Functions Γ, Φ, and Λ estimate aspect similarity by three different criteria, bounded in [0,1].

Semantic Similarity (Γ): Compares two aspect co-occurrence vectors to estimate their similarity [12]. The co-occurrents of an aspect a_i are the other aspects that have a first order co-occurrence with a_i within a sentence window. By passing this window over the entire corpus we obtain, for each aspect a_i, a vector of its co-occurrent aspects. The vector of co-occurrent aspects represent the global context of the aspect with respect to the other aspects in $\mathcal{A}$, and we use it to estimate the semantic similarity between aspects. That is to say, we consider that two aspects are semantically close if the co-occurrence vectors of both aspects, with respect to all other aspects in $\mathcal{A}$, are similar. Figure 2 shows the co-occurrent relations between some of the top most frequent aspects in the aspect vocabulary of DSLR cameras, where the size of the nodes represent the frequency of occurrence of the aspects over the reviews of the DSLR corpus (bigger nodes are more frequent), and the edges the strength of the co-occurrence between aspects (the wider the edge the more times those aspects co-occur in the same sentence). Both the aspect vocabularies and the corpora will be introduced later in Section 5.

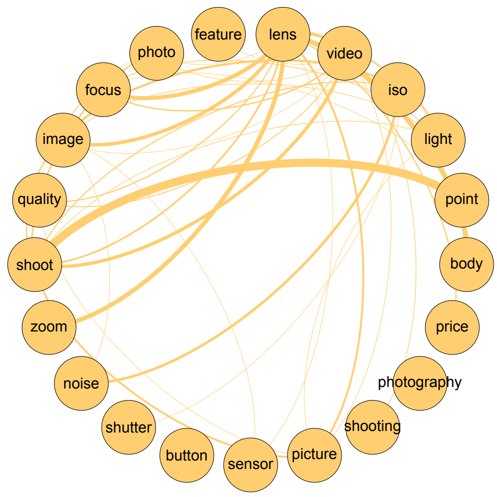

Fig. 2. Co-occurrences between pairs of the most frequent aspects of the aspect vocabulary of DSLR cameras.

String Similarity (Φ): Uses the Jaro-Winkler distance [14] to estimate the string similarity between two aspects. This string similarity compares the characters of two strings giving more importance to the left-most characters of the words (to boost aspects with similar lemmas).

Taxonomic Similarity (Λ): *PhotoDict* is a small taxonomy of camera-related terms, where similarity is measured as the shortest path between two terms. We automatically generated *PhotoDict* from a camera related website [4].

From aspect similarity we define the similarity Sim_G between two groups of aspects G_n and G_m:

$$Sim_G(G_n,G_m) = \frac{1}{|G_n||G_m|}\sum_{i=1}^{|G_n|}\sum_{j=1}^{|G_m|} Sim_A(a_i,a_j)$$

There is a special treatment of compound nouns in clustering. Since compound nouns are formed by two or more words (e.g. *image quality*), we group them with the most frequent aspect in the compound.

The result of the hierarchical clustering is a dendrogram (or clustering tree) of aspects; Figure 3 shows a portion of the resulting dendrogram of the DSLR aspect vocabulary (see Section 5), considering only a representative subset.

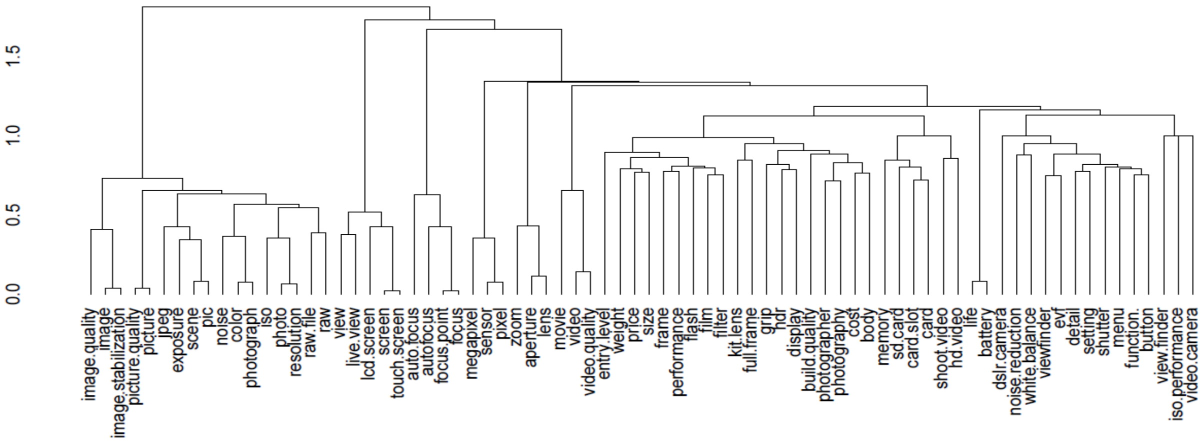

Fig. 3. Representative portion of the dendrogram for the DSLR cameras, from which the concept vocabulary for DSLR cameras is created.

Since hierarchical clustering gives multiple partitions over the initial elements at different levels, we have to select a single partition over $\mathcal{A}$ to create a concept vocabulary $\mathcal{C}$.

2.2. Concept Discovery

We are interested in selecting a partition from the hierarchical clustering dendrogram such that that the parts constitute basic level concepts (BLC) of digital cameras. This selection should be guided by the usage of the aspects occurring in a corpus of reviews $\mathcal{K}$. The selected partition will become our concept vocabulary $\mathcal{C}$ by assigning every aspect in $\mathcal{A}$ to a concept in $\mathcal{C}$. Therefore, every concept in $\mathcal{C}$ is formed by a set of aspects that are, in their usage, coherent around a basic level concept.

We will consider that an aspect group G is a good candidate for being a BLC when, for all aspects in G, the average polarity of the sentences related to each aspect cohere with respect to each product.

To select the best partition, we cut the dendrogram at different levels. Then, for each partition Π, we analyze the coherence degree of the sentiment values in each aspect group in Π. If the sentiments of the aspects of a group G cohere into a clear positive, negative, or neutral value, we consider G a potential basic level concept. For instance, let *picture*, *photo* and *image* be three aspects in a group. If those three aspects are used by people to refer to the same concept ('picture obtained by my digital camera'), then the sentiment values of those aspects with respect to the reviews of each product should have a high coherence degree.

Thus, we define the *Partition Ranking* score $R(\Pi)$ of a partition Π by aggregating the sentiment coherence of the aspect groups in Π, as follows:

$$R(\Pi) = \frac{1}{|\Pi|} \sum_{G_i \in \Pi} IS(G_i)$$

where $G_i \in \Pi$ and $|\Pi|$ is the number of aspect groups in Π. The higher $R(\Pi)$, the better the partition Π.

$IS(G)$ estimates the coherence degree of an aspect group G as the average sentiment similarity among the aspects in G using the average cosine similarity:

$$IS(G) = \frac{1}{|G| \cdot (|G|-1)} \sum_{i=1}^{|G|} \sum_{j=1, j \neq i}^{|G|} cos(D(a_i), D(a_j))$$

where $cos(D(a_i), D(a_j))$ is the cosine of the angle between aspect vectors $D(a_i)$ and $D(a_j)$. Finally, an aspect vector $D(a)$ is:

$$D(a) = (S_{av}(p_i, a))_{i \in 1, \ldots, |\mathcal{P}|}$$

where $p_1, \ldots, p_{|\mathcal{P}|}$ is the set of products $\mathcal{P}$ described in a review corpus $\mathcal{K}$ and $S_{av}(p_i, a) \in [0, 1]$ is the normalized sentiment average over the set of sentences from the reviews of product p_i in which aspect a occurs. Therefore, the aspect vector $D(a)$ contains the average polarity of aspect a, considering all sentences from user reviews over $\mathcal{P}$. Notice that, by comparing two aspect vectors $D(a_i)$ and $D(a_j)$, we can assess the polarity coherence between two aspects a_i and a_j over the set of products of a corpus $\mathcal{K}$.

The concept vocabulary $\mathcal{C}$ we took corresponds to the partition with greater $R(\Pi)$ (in our experiments we considered only partitions with 35 to 45 groups, a reasonable concept vocabulary size).

Concept	Aspects in Concept
Storage	storage, capacity, sd card, sdhc card, cf card
Button	lag, shutter release, shutter speed, shutter lag, shutter button, button, button layout
Battery	battery, battery life, battery pack

Table 1

Three example basic level concepts and their group of aspects.

3. Bundle of Arguments

In this section we characterize the a product by a bundle of arguments that use concepts in $\mathcal{C}$.

Let $p \in \mathcal{P}$ be a product described in the reviews of a corpus $\mathcal{K}$, and $C \in \mathcal{C}$ a concept. We define $Occ(p,C)$ as the set of sentences from the reviews that describe individual experiences with product p in which any of the aspects that form the concept C appears. By analyzing the sentiment values of $Occ(p,C)$, we infer whether the people's experiences about a concept C of a product p have a positive or negative overall sentiment. If the overall polarity of the occurrences of C over the reviews of p is positive, we consider C to be a *pro* argument (Arg^+) of p. If the overall polarity is negative, we consider C a *con* argument (Arg^-) of p. Finally, if the overall polarity is not clearly positive or negative, we consider C a *moot* argument (Arg^0). An argument:

$$Arg = \langle p, C, s \rangle$$

is formed by a product $p \in \mathcal{P}$, a concept $C \in \mathcal{C}$ and an aggregated sentiment s calculated by aggregating the sentiment values of $Occ(p,C)$. Pro, con, and moot arguments are defined as follows:

$$Arg = \begin{cases} Arg^+ & \text{if } Arg.s > \delta \\ Arg^- & \text{if } Arg.s < -\delta \\ Arg^{\emptyset} & \text{if } -\delta \leq Arg.s \leq \delta \end{cases} \tag{1}$$

where δ is a threshold that determines when arguments are considered clearly pros or cons —and moot otherwise. We will see later that δ depends on the bundle type ($\delta_G, \delta_\sigma, \delta_F$).

By considering the pro, con and moot arguments of a product p over the set of all concepts in $\mathcal{C}$, we obtain a characterization about what people like or dislike about p. The union of all these arguments form the bundle of arguments B of a product p:

$$B(p) = \bigcup_{C \in C} Arg\,\langle p, C, s \rangle$$

In this work we consider three different methods to create a bundle of arguments: Gini (B_G), Agreement (B_σ), and Cardinality (B_F) bundles. Each bundle type (B_G, B_σ and B_F) is built by a different sentiment aggregation measure. As we will see later, the parameter Δ defines as moot those arguments with a very small $Occ(p,C)$.

3.1. Gini Bundle of Arguments (B_G)

An argument in B_G has the form $\langle p, C, S_G(p,C) \rangle$, where the polarity value S_G is calculated using the average sentiment $S_{av}(p,C)$ together with the *Gini Coefficient* [17] to penalize it according to the degree of dispersion of sentiment values:

$$S_G(p,C) = \begin{cases} 0 & \text{if } |Occ(p,C)| < \Delta \\ & \text{or } -\delta_G < S(p,C) < \delta_G, \\ S(p,C) & \text{otherwise.} \end{cases}$$

where $S(p,C) = S_{av}(p,C)(1 - Gini(p,C))$.

Notice that, when $|Occ(p,C)| < \Delta$, we consider that we do not have enough reviews of product p with concept C and we assign a neutral sentiment value. Similarly, when $-\delta_G < S_{av}(p,C) \cdot (1 - Gini(p,C)) < \delta_G$, we consider that the polarity is not strong enough to define an argument as a pro or a con, and we assign a neutral sentiment value (recall Equation 1). δ_G is set to 0.1 in the experiments.

3.2. Agreement Bundle of Arguments (B_σ)

Let $Dev(p,C)$ be the standard deviation of the sentiment values of $Occ(p,C)$. The agreement sentiment measure $S_\sigma(p,C)$ is the sentiment average of the sentiment values of the sentences in $Occ(p,C)$, for those concepts whose $Dev(p,C) < \delta_{max}$. This measure uses two threshold parameters δ_{max} and δ_σ. First, δ_{max} specifies the maximum acceptable standard deviation over the distribution of sentiment values in $Occ(p,C)$: when $Dev(p,C) > \delta_{max}$ we consider that we have no grounds for an informed decision on the overall polarity of C with respect to product p. Second, δ_σ specifies the threshold for an argument sentiment value to be considered pro, con, or moot argument (see Equation 1).

An argument in B_σ has the form $\langle p, C, S_\sigma(p,C) \rangle$, where S_σ is defined as follows:

$$S_\sigma(p,C) = \begin{cases} 0 & \text{if } Dev(p,C) > \delta_{max} \\ & \text{or } |Occ(p,C)| < \Delta, \\ S_{av}(p,C) & \text{otherwise.} \end{cases}$$

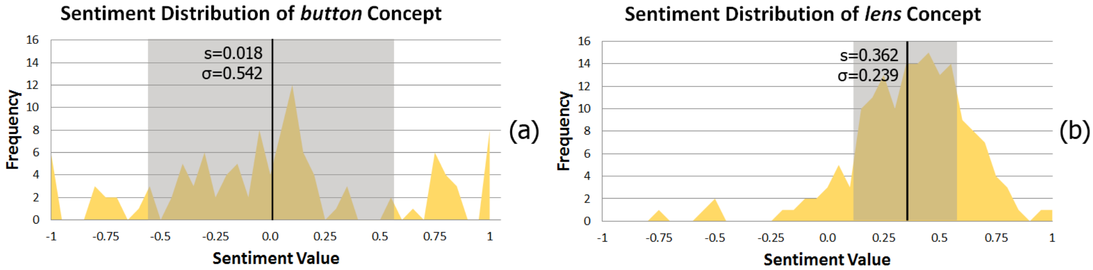

Fig. 4. Sentiment value distribution (a) *button* concept and (b) *lens* concept for Pentax K-5. Notice that values have a higher degree of dispersion in (a) than in (b). Furthermore, sentiment values for *lens* concept in Pentax K-5 (b) are clearly positive, with an average sentiment $s = 0.362$.

Parameter δ_σ is set to 0.1 in the experiments. Similarly as before, when $|Occ(p,C)| < \Delta$ we consider that we do not have enough reviews of product p with concept C and we assign a neutral sentiment value.

Figure 4 presents the sentiment value distribution of two arguments of Pentax K-5, *button (a)* and *lens (b)*. The *button* argument of the Pentax K-5 has a sentiment value deviation $\sigma = 0.542$, showing a high dispersion of sentiment values for concept *button* among the reviews of Pentax K-5. Since the deviation of the sentiment values of *button* is higher than δ_{max}, we have no clear overall polarity. On the other hand, the deviation of the sentiment values of *lens* is lower than δ_{max} and has a positive average sentiment ($0.235 > \delta_\sigma$). Therefore, concept *lens* is considered part of a pro argument with respect to Pentax K-5.

3.3. Cardinality Bundle of Arguments (B_F)

The cardinality bundle is created by comparing the number of positive (O^+) versus negative (O^-) occurrences of a concept C in $Occ(p,C)$.

$$O^+(p,C) = |\{x \in Occ(p,C) | s(C,x) > 0\}|$$
$$O^-(p,C) = |\{x \in Occ(p,C) | s(C,x) < 0\}|$$

where $s(C,x)$ is the sentiment value in $[-1,1]$ of concept C in sentence $x \in Occ(p,C)$. The comparison of positive versus negative number of occurrences of concept C in the reviews of product p is the function $O(p,C) \in [-1,1]$:

$$O(p,C) = \left(2 \cdot \frac{O^+(p,C)}{O^+(p,C) + O^-(p,C)}\right) - 1$$

Thus, an argument in B_F has the form $\langle p, C, S_F(p,C)\rangle$ where S_F is:

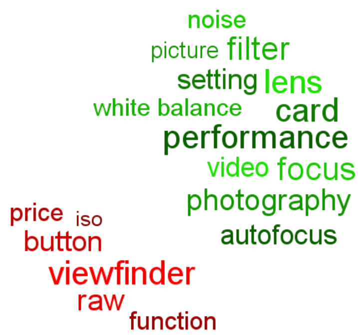

Fig. 5. Gini Argument Bundle B_G of the Canon EOS Rebel T4i camera. Pros are in green and above and cons in red and below. Letter size corresponds to the strength of the argument sentiment, larger means stronger polarity in sentiment value.

$$S_F(p,C) = \begin{cases} 0 & \text{if } O(p,C) = 0 \\ & \text{or } |Occ(p,C)| < \Delta, \\ O(p,C) & \text{otherwise.} \end{cases}$$

Notice that $S_F(p,C)$ takes values on $(0,1]$ if $O^+ > O^-$, and in $[-1,0)$ if $O^+ < O^-$. Also, when $|Occ(p,C)| < \Delta$ we consider that we do not have enough occurrences of product's p concept C to make an informed decision, and we assign a neutral value to the sentiment of the argument. In the experiments we use $\delta_F = 0$ (recall Equation 1) as the threshold that determines if a cardinality argument is a pro, con or moot.

Figure 5 shows the pro and con arguments of the Gini bundle of arguments (B_G) of the Canon EOS Rebel T4i (moot arguments are not represented). Each word represents an argument: pros in green and above, and cons in red and below.

As a final step in order to create a collection of bundles (B_G, B_σ, B_F) for each product in $\mathcal{P}$, we first need, considering the whole $\mathcal{P}$ in a corpus $\mathcal{K}$, to rescale the sentiment values of the arguments that form the bundles of the collection of bundles in a way that the most positive argument sentiment about a concept has a sentiment 1, and the most negative a sentiment -1. We rescale the rest of the sentiment values accordingly (see [4] for more details). This way, considering a collection of product bundles of a given type, the product with the best sentiment over a concept has a sentiment value of 1. When all arguments of a bundle $B(p)$ are rescaled we call it a normalized bundle $\overline{B}(p)$.

4. User Query over Product Bundles

A user query defines the preferences of a user expressed using the concept vocabulary $\mathcal{C}$. Since not all preferences are equally important for the user, every preference over a concept has a utility value. Given the products $\mathcal{P}$ characterized with the normalized bundles of arguments $\overline{B}(p)$, we need to determine which $p \in \mathcal{P}$ has a higher level of query satisfaction.

We define a *user query* $Q = \{(C_j, U(C_j))\}_{j=1,\dots,k}$ with $k \leq |\mathcal{C}|$. Each concept utility pair $(C_j, U(C_j))$ expresses a preference of user over a concept C_j and its strength with a utility degree $U(C_j) \in [0.5, 1]$.

For instance in a query $Q = \{(lens, 0.9), (video, 0.6)\}$, the user prefers high quality lens and video, although the quality of the lens is more important than that of the video. Furthermore, lens and video are more important than any other arguments characterizing the camera.

We will now define the *Degree of Query Satisfaction*, $DS(Q, \overline{B})$), that determines the degree to which a normalized bundle $\overline{B}$ satisfies a user query Q, using the framework of fuzzy logic. Since t-norms and implications in fuzzy logic are defined in the interval $[0, 1]$ [5], we need to rescale the sentiment values of all arguments that form all normalized product bundles from [-1,1] to [0,1]. We do so by applying the linear mapping $\widehat{s} = \frac{s+1}{2}$. For example, consider a normalized argument $\langle p$, *lens*, $0.83\rangle \in \overline{B}(p)$, the sentiment of the rescaled argument will be $\widehat{s} = 0.915$, and the resulting rescaled argument is $\langle p$, *lens*, $0.915\rangle \in \widehat{B}(p)$; the neutral value 0 in [-1,1] is mapped to the neutral value 0.5 in [0,1].

The degree of query satisfaction is defined by aggregating the degrees to which an argument, with respect to concept C, satisfies a user preference with respect to the same concept C. Therefore, we first define a concept-wise satisfaction degree, using the notion of fuzzy implication associated to the t-norm product ($\Rightarrow_\otimes$). The fuzzy implication $U(C_j) \Rightarrow_\otimes \widehat{s_j}$ models this notion of degree of satisfaction: if the sentiment $\widehat{s_j}$ of an argument related with concept C_j is higher than the user preference $U(C_j)$, then the user preference is satisfied (and satisfaction degree is 1). On the other hand, if the sentiment $\widehat{s_j}$ provided by an argument is lower than the user preference $U(C_j)$, then the user preference is satisfied to a degree lower than 1:

Q_1 Preferences	$(C_1, 0.7)$	$(C_2, 0.6)$	*DS*
$\widehat{B}_F$(D7100)	0.75	1.00	
$\widehat{B}_F$(EOS70D)	0.97	0.50	
$U(C_j)$ for $\widehat{B}_F$(D7100)	1.00	1.00	**1.00**
$U(C_j)$ for $\widehat{B}_F$(EOS70D)	1.00	0.83	0.83

Q_2 Preferences	$(C_1, 0.7)$	$(C_2, 0.6)$	$(C_3, 0.9)$	*DS*
$\widehat{B}_F$(D7100)	0.75	1.00	0.64	
$\widehat{B}_F$(EOS70D)	0.97	0.50	1.00	
$U(C_j)$ for $\widehat{B}_F$(D7100)	1.00	1.00	0.72	0.72
$U(C_j)$ for $\widehat{B}_F$(EOS70D)	1.00	0.83	1.00	**0.83**

Table 2

Degree of satisfaction of two cameras for each individual preference and the overall *DS* for the query Q_1 and the Q_2, where C_1 = *picture*, C_2 = *resolution*, C_3 = *video*, and *DS* is the degree of Query Satisfaction.

$$U(C_j) \Rightarrow_\otimes \widehat{s_j} = \begin{cases} 1 & \text{if } U(C_j) \leq \widehat{s_j}, \\ \frac{\widehat{s_j}}{U(C_j)} & \text{otherwise.} \end{cases} \quad (2)$$

where $\widehat{s_j}$ is the rescaled sentiment value of argument $\langle p, C_j, \widehat{s_j}\rangle$ and $\frac{\widehat{s_j}}{U(C_j)}$ is the satisfaction degree.

Given a query Q with k preferences we can now infer k concept-wise satisfaction degrees with respect to a bundle $\widehat{B}(p)$. We need now to aggregate these k satisfaction degrees into an overall degree of bundle satisfaction ($DS(Q, \widehat{B}(p))$). We do so using the conjunction of these resulting concept-wise satisfaction degrees. Since conjunction in fuzzy logic are represented by t-norms, we use the product t-norm (in consonance with the product implication used in Eq. 2):

$$DS(Q, \widehat{B}(p)) = \prod_{j=1}^{k} (U(C_j) \Rightarrow_\otimes \widehat{s_j})$$

where $\widehat{s_j}$ is the rescaled sentiment value of argument $\langle p, C_j, \widehat{s_j}\rangle$ of the argument bundle $\widehat{B}(p)$, and $\widehat{B}$ is a rescaled argument bundle (either $\widehat{B}_G$, $\widehat{B}_\sigma$ or $\widehat{B}_F$).

Table 2 shows the degree of satisfaction of two user queries Q_1 and Q_2 against the cardinality bundles of two cameras: Nikon D7100 and Canon EOS70D (sentiment values are rescaled). The first query is created by a user who likes to go hiking and is looking for a camera to capture landscape and nature while valu-

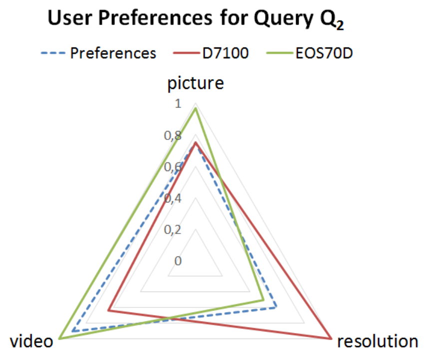

Fig. 6. Value-triangle of user preferences for the three concepts *picture*, *resolution* and *video* compared with the value-triangles of bundle $\widehat{B}_F(\text{EOS70D})$ and bundle $\widehat{B}_F(\text{D7100})$.

ing fine detail. Assume her query is $Q_1 = \{(picture,\ 0.7), (resolution,\ 0.6)\}$ because she wants a camera with good image quality and resolution. Table 2 shows on the first two rows the argument sentiment values of the two cameras corresponding to the concepts appearing in the query. The second two rows show the satisfaction degree of the two cameras for each preference and the third row overall *DS* for the query. Notice that satisfaction is 1 when the argument sentiment value is higher than the query's utility value for that concept.

The second half of Table 2 shows another query, $Q_2 = \{(picture,\ 0.7), (resolution,\ 0.6), (video,\ 0.9)\}$. Q_2 is created by a user that, besides hiking, also loves recording video. Figure 6 shows a comparison between the user preferences regarding concepts 'picture', 'video' and 'resolution', and the sentiment of the corresponding arguments of the two camera bundles, $\widehat{B}_F(\text{D7100})$ and $\widehat{B}_F(\text{EOS70D})$. Now, according to user reviews, Canon EOS70D has an outstanding video quality (1.0), while Nikon D7100 has an average quality video (0.64). Because of this newly added preference now the higher ranking camera for Q_2 is Canon EOS70D instead of Nikon D7100, the best ranking camera for Q_1. This is clear when we observe Figure 6, where EOS70D value-triangle is closer to the user preferences value-triangle.

5. Evaluation

We evaluate the argument bundles of products from three different digital camera categories extracted from Amazon: Digital SLR, Compact, and Point & Shoot cameras. For this purpose we compare the bundles of arguments we create in the three categories with a set of bundles created from DPReview.com, a renowned website specialized in digital cameras. Moreover, we are keen to study if there are significant differences between the sets of pros and cons depending the three types of bundles of arguments, B_G, B_σ and B_F, and for the three camera categories. We also assess the impact that the number of reviews of each product has over the quality of their argument bundles. Furthermore, we evaluate the precision and recall of the product bundles by comparing them with the expert evaluations of products presented in DPReview. Finally, we present a ranking strategy for product bundles and compare the rankings of products obtained with each bundle type (B_G, B_σ, B_F) with two external product rankings (those of DPReview and Amazon).

This section is structured as follows: Section 5.1 introduces the three camera corpora and the corresponding aspect vocabularies used to evaluate this experiment: The DSLR corpus $\mathcal{K}_D$, the Compact corpus $\mathcal{K}_D$, and the Point & Shoot $\mathcal{K}_D$. Then, in Section 5.2, we create three concept vocabularies ($\mathcal{C}_D$, $\mathcal{C}_C$ and $\mathcal{C}_P$) for the tree corpora (as described in Section 2). Section 5.3 explains how we create and compare the bundles of arguments of the cameras in the corpora $\mathcal{K}_D$, $\mathcal{K}_C$ and $\mathcal{K}_P$. Section 5.4 evaluates the bundles of arguments of the DSLR cameras by comparing them with those of DPReview. Finally, in Section 5.5 we compare the rankings of products obtained with each bundle type with the rankings extracted from DPReview and Amazon.com.

5.1. *Review Corpora and Aspect Vocabularies*

During September 2015, we extracted more than 100,000 reviews from `Amazon.com` corresponding to 2,264 digital cameras from three different categories: Digital SLR, Compact System Cameras, and Point & Shoot. We filtered out those products that were older than January 1st 2010 and had less than 15 different user-generated reviews. Then, we united all synonymous products leaving us data for 102 products in the DSLR category, 95 in Compact category, and 599 products in Point & Shoot category. Finally, we grouped the resulting product reviews in three corpora: $\mathcal{K}_D$ for DSLR cameras, $\mathcal{K}_C$ for Compact cameras, and $\mathcal{K}_P$ for Point & Shoot cameras. Each corpus is formed by a set of product-reviews pairs $(p_i, Rev(p_i))$, where p_i is a digital camera and $Rev(p_i)$ is the set of reviews about camera p_i. Table 3 shows the quantity of products and reviews on each corpus. Notice that the num-

Category	$\mathcal{K}_D$	$\mathcal{K}_C$	$\mathcal{K}_P$
No. of Products	102	95	599
No. of Reviews	7,552	6,334	84,138
Avg. Reviews/Product	74.03	66.67	140.46

Table 3

The three corpora: DSLR ($\mathcal{K}_D$), Compact ($\mathcal{K}_C$), and P&S ($\mathcal{K}_P$).

Concept Vocabularies	#Concepts	#Aspects
$\mathcal{C}_D$	41	225
$\mathcal{C}_C$	39	197
$\mathcal{C}_P$	39	179

Table 4

Number of concepts and aspects for each concept vocabulary.

Top 10 in $\mathcal{C}_D$	Top 10 in $\mathcal{C}_C$	Top 10 in $\mathcal{C}_P$
picture	image	picture
lens	lens	zoom
video	price	video
screen	video	price
focus	focus	battery
button	zoom	screen
iso	button	price
photography	screen	focus
price	battery	button
battery	sensor	flash

Table 5

Top 10 most frequent concepts of $\mathcal{C}_D$, $\mathcal{C}_C$, and $\mathcal{C}_P$.

ber of products (and reviews) in $\mathcal{K}_P$ is much larger than those in $\mathcal{K}_D$ and $\mathcal{K}_C$.

For each of the three corpora of digital cameras, we created a corresponding aspect vocabulary, as described in our previous work on social recommender systems [15] (see also Section 2). Thus, we created three distinct aspect vocabularies $\mathcal{A}_D$, $\mathcal{A}_C$, and $\mathcal{A}_P$, corresponding to the corpora $\mathcal{K}_D$, $\mathcal{K}_C$, and $\mathcal{K}_P$. The three aspect vocabularies consist of set of salient aspects used to define important characteristics of photographic digital cameras. Moreover, as we show in [4], the three aspect vocabularies contain different sets of aspects.

5.2. *Creation of the Concept Vocabularies*

For each aspect vocabulary ($\mathcal{A}_D$, $\mathcal{A}_C$ and $\mathcal{A}_P$), we create a hierarchical clustering dendrogram as explained in Section 2.2 (one for each camera category). Then, we select the partition with highest $R(\Pi)$ from each dendrogram, removing those groups $G_i \in \Pi$ whose $|Occ(\mathcal{P}, G_i)| \leq 100$, and only considering partitions with 35 to 45 groups, a reasonable concept vocabulary size for our purposes. The selected partition Π consists of a collection of groups of aspects, that are considered the collection of basic level concept (BLC) for that corpus. These sets of BLCs form the concept vocabularies, $\mathcal{C}_D$, $\mathcal{C}_C$ and $\mathcal{C}_P$, for each camera category. Each vocabulary determines the set of concepts that we use to interpret the reviews from each corpus.

Table 4 shows the quantity of concepts and aspects that form the three concept vocabularies: $\mathcal{C}_D$ (with 41 concepts), $\mathcal{C}_C$ (with 39 concepts), and $\mathcal{C}_P$ (with 40 concepts). When creating the concept vocabularies, we discarded those aspect groups from the selected partition that had less than 100 occurrences in the reviews of a corpus. We interpret the low number of occurrence of the aspects within those concepts as an indicator that they are not deemed important by people when describing their experiences with digital cameras.

Table 5 shows the top 10 most frequent concepts in $\mathcal{C}_D$, $\mathcal{C}_C$, and $\mathcal{C}_P$, where the name of a concept corresponds to the most frequent aspect included in that concept. These 10 more frequent concepts are similar in the three camera corpora, with a few exceptions such as concept 'zoom', and 'iso' (deemed important for DSLR cameras, but not for Compact and Point & Shoot cameras). However, the concepts with lower frequency are less similar among the vocabularies. The difference comes not only from the name of the concept but from the aspects grouped under a concept in different vocabularies. For instance DSLR and Compact differentiate some picture-related concepts such as 'iso', 'noise' or 'resolution', while in Point & Shoot they are conflated in the concept 'picture' [4]. A similar situation is observed in concept 'button': this concept groups aspects 'button shutter', 'button' and 'shutter' in the three concept vocabularies $\mathcal{C}_D$, $\mathcal{C}_C$ and $\mathcal{C}_P$. However, the aspects 'button layout' and 'release shutter' are only found in concept 'button' of $\mathcal{C}_D$, and not in concept 'button' of $\mathcal{C}_C$ and $\mathcal{C}_P$; while aspects 'menu button', 'lag' and 'shutter lag' are only found in concept 'button' of $\mathcal{C}_C$. Furthermore, some concepts only exist in one of the three vocabularies, e.g. 'waterproof' is only present in $\mathcal{C}_P$.

5.3. *Comparing Argument Bundles Types*

For each product of the corpora $\mathcal{K}_D$, $\mathcal{K}_C$ and $\mathcal{K}_P$, we create three argument bundle types (B_G, B_σ and B_F). As described in Section 3, the argument bundles of the products of a corpus $\mathcal{K}_x$ are expressed with the cor-

	Gini Bundle B_G	Agreement Bundle B_σ	Cardinality Bundle B_F
DSLR ($\mathcal{K}_D$)			
Avg # pros	11.82	15.61	16.06
Avg # cons	0.59	4.08	3.19
Avg # moots	28.59	21.31	21.75
Compact ($\mathcal{K}_C$)			
Avg # pros	9.42	12.70	13.16
Avg # cons	0.47	3.66	2.90
Avg # moots	29.11	22.64	22.94
Point & Shoot ($\mathcal{K}_P$)			
Avg # pros	9.64	18.49	19.01
Avg # cons	0.82	6.59	5.58
Avg # moots	28.54	13.92	14.41

Table 6

Average number of pro, con and moot arguments for the three bundle types B_G, B_σ, and B_F for $\mathcal{K}_D$, $\mathcal{K}_C$, and $\mathcal{K}_P$.

responding concept vocabulary $\mathcal{C}_x$. So, the argument bundles of the products in $\mathcal{K}_D$ are created using the DSLR concept vocabulary $\mathcal{C}_D$.

In this section we study the differences between the three bundle types B_G, B_σ and B_F for the products of each camera corpus $\mathcal{K}_D$, $\mathcal{K}_C$, and $\mathcal{K}_P$. Since the criteria to establish an argument as pro, con, or moot varies between the three bundle types, the quantity of pros, cons, and moot arguments obtained by each bundle type may differ. Table 6 presents a comparison between the average quantity of pro, con and moot arguments of each bundle type for DSLR cameras $\mathcal{K}_D$, Compact cameras $\mathcal{K}_C$, and Point & Shoot cameras $\mathcal{K}_P$. Notice that the values presented in the table are averages over all products of the same corpus. As we will show later, products with more reviews usually have more pros and cons —and much less moots.

In Table 6, the Agreement (B_σ) and Cardinality (B_F) bundles have a similar average number of pros and cons, while Gini (B_G) bundles have slightly lower quantity of pros and cons for the three camera corpora (DSLR, Compact and Point & Shoot). The Gini average tends to move the argument sentiment value towards 0 when there is dispersion in the distribution of sentiment values, and thus more arguments tend to be moots. The highest number of pros in a bundle type is obtained with the cardinality bundle B_F of the Point & Shoot corpus $\mathcal{K}_P$, while the least number of pros is obtained by the Gini bundle B_G of the Compact corpus $\mathcal{K}_C$.

On the other hand, the quantity of pros is higher than the quantity of cons for all bundle types (B_G, B_σ, and B_F) for the three corpora. Either the SmartSA sentiment analysis system is biased towards positive sentiments, or the reviews of our three corpora contain more positive than negative sentences referencing the concepts of the concept vocabularies. Notice that the quantity of pro and con arguments is directly related to the average quantity of reviews per product, presented in Table 3. For instance, the bundles of arguments of the Point & Shoot corpus $\mathcal{K}_P$, with an average quantity of reviews per product of 140.46, contain more pro and con arguments than the bundles of the other two corpora $\mathcal{K}_C$ and $\mathcal{K}_P$, with an average quantity of reviews per product of 74.03 and 66.67, respectively.

Next we analyze which concepts are found in the pro arguments of the three bundle types for each product. Figure 7 shows, for each product in the horizontal axis, the quantity of concepts in pro arguments that are present in 1, 2 or 3 bundle types. The left vertical axis (*#Concepts*) shows that most pros (almost 8 out of 10) are present in 2 or 3 bundle types of a product, a good indicator of the consistency of our approach. This means that a pro argument concept in a B_G is also likely to be present in a pro argument in B_σ, or B_F, or both. The right vertical axis (*#Pro concepts occurrences*) shows the number of concept occurrences in the reviews of each product, displayed as a triangle in Fig. 7. The number of pro arguments in a product bundle is directly related to the number of concept occurrences in the reviews of that product. Similar results are obtained when analyzing concepts in con arguments.

5.4. Bundle of Arguments Evaluation

To evaluate the quality of bundles, we compared the product pros and cons textual descriptions we found on the DPReview website with the bundles of arguments of the 15 products of the DSLR cameras $\mathcal{K}_D$ with the highest number of reviews. The DPReview pros and cons of a product are in two separate lists with item texts such as ‘good detail and color in JPEGs at base ISO (pro)’ or ‘buggy Live View / Movie Mode (con)’. In order to compare the DPReview pro and con items with our bundles of arguments, we first manually identify the concepts, from our DSLR vocabulary $\mathcal{C}_D$, referenced in each item text. For instance, we consider that the DPReview sentence ‘good detail and color in JPEGs at base ISO’ refers positively to the concepts ‘jpeg’, ‘color’ and ‘picture’, whilst ‘buggy Live View / Movie Mode’ refers negatively to concepts ‘live view’ and ‘video’. Those sentences from DPReview that did not clearly refer to a concept in $\mathcal{C}_D$ were ignored. By grouping the vocabulary concepts present in the DPRe-

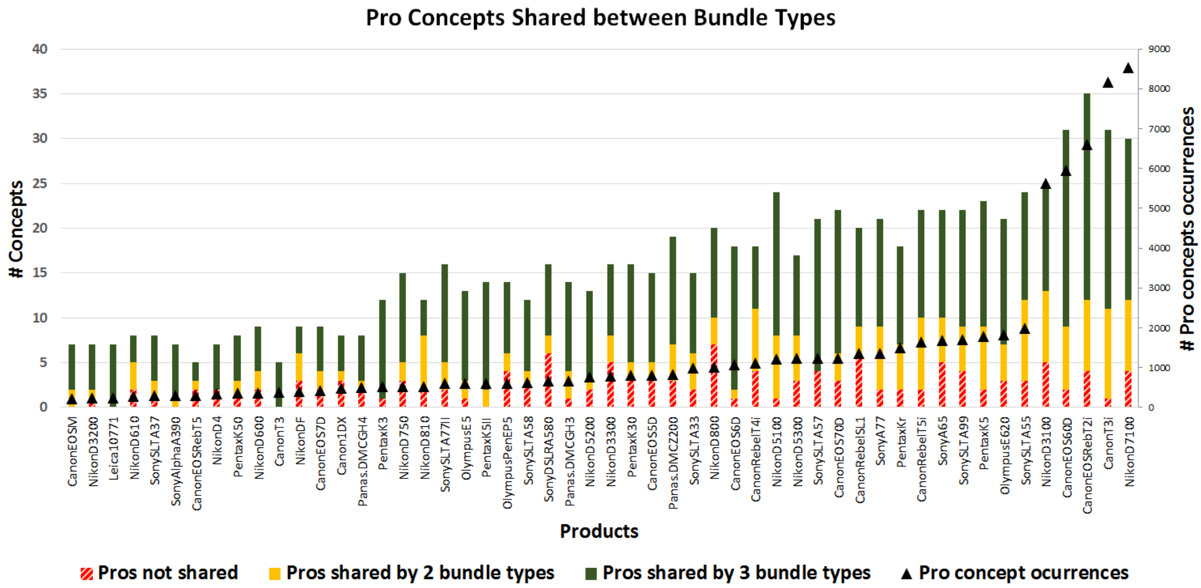

Fig. 7. Quantity of pros shared between the three bundles of arguments B_G, B_σ and B_F of the top 50 products with more reviews of DSLR corpus $\mathcal{K}_D$, together with the number of occurrences of the pro concepts in the reviews of the product.

view pro and con items of a product, we create the sets of DPReview pro and con arguments ($Pros_{dp}$ and $Cons_{dp}$), but without associating any numeric value of sentiment —only the positive or negative nature of the polarity expressed in the text. We compare those DPReview $Pros_{dp}$ and $Cons_{dp}$ sets with the pros and cons of the three different bundles of arguments of each product —without taking into account the sentiment values, only whether the concept is in a pro or in a con.

Table 7 presents the average precision, recall and F_2-score between the sets of pros and cons of the three bundle types of the 15 DSLR cameras with more reviews and those of DPReview. We use the F_2-score to weight recall higher than precision, since we are keen to study whether the three different bundle types identify as pros and cons the same concepts listed in DPReview. Furthermore, we analyze the percentage of *contradictions*, which are those concepts selected as pros in our bundles of arguments but considered cons in DPReview or vice versa. A low rate of contradictions is a good indicator of the quality of the bundles if we take DPReview as a standard for comparison.

The argument bundle type that performs best for the pro arguments of the selected DSLR products shown in Table 7 is the cardinality bundle B_F, with an average recall of 0.822 and an F_2-score of 0.733. This means that the 82.2% of the arguments listed as pros of product a p in DPReview also form part of the pros of the cardinality bundle $B_F(p)$. On the other hand,

		Precision	Recall	F_2-score	*Contr*
Pros	B_G	**0.567**	0.644	0.627	**0.004**
	B_σ	0.506	0.761	0.691	0.135
	B_F	0.513	**0.822**	**0.733**	0.065
Cons	B_G	0.333	0.046	0.056	0.046
	B_σ	0.285	0.558	0.468	0.132
	B_F	0.388	0.488	0.464	0.165

Table 7

Measure on precision, recall, F_2-score and contradictions between pros and cons of bundles B_G, B_σ and B_F, of DSLR products from corpus $\mathcal{K}_D$, with respect to DPReview pros and cons.

the sets of cons of all three bundles of arguments perform poorly. The reason is the difference in granularity between our concept vocabulary and those concepts used in DPReview. For us, the granularity level is given by our concept vocabulary, while DPReview sentences address concepts that are at different levels of granularity. Furthermore, the granularity of DPReview sentences varies whether the sentence is a pro or a con. Pro sentences in DPReview tend to be more general: ‘camera buttons and dials are useful and easily configurable’, while con sentences tend to be more specific: ‘the video dial is not easily accessible’. Although for us both sentences reference the same concept (‘button’ in this example), the DPReview pro sentence addresses a more general view of the buttons of the camera than the con sentence.

Furthermore, the precision values of the 3 bundle types are lower than 0.6, suggesting that the sets of pros of the bundles of arguments are richer in concepts compared to the lists of DPReview. This is not surprising, since the sets of DPReview pros and cons are not exhaustive but a short list of the concepts that stand out from their point of view. The average quantity of pro arguments in a bundle is 12-14, while the average pro set size of DPReview identified arguments is 7-9.

Finally, notice the number of contradictions between the bundles of arguments and the DPReview sets is low. Nevertheless, we are interested in studying which concepts occur more often in contradictions.

The most common contradictions between the bundles and the set of pro and con extracted from DPReview for the 15 selected products are: 'battery' (10), 'viewfinder' (5), 'recording' (5) and 'button' (3). In DPReview, 'battery' is often selected as a pro, however it is usually present in con arguments in our bundles. That is because in the reviews people usually complain about the battery of a camera, while they do not seem to express positive opinions on cameras with a good battery (it would seem it is taken as a given). Other frequent contradictions are 'viewfinder', 'recording' and 'button'. This is because in DPReview those are commonly selected as cons for having suboptimal behavior in certain types of situations (e.g. 'the video dial is not easily accessible') while the overall opinions about the rest of the buttons are positive. Therefore, our bundles will capture this average higher granularity sentiment of 'button'. Similar situations are observed for 'recording' and 'viewfinder' concepts.

5.5. *Bundle Evaluation by Product Ranking*

In this section we are interested in comparing the bundles of arguments with the camera descriptions of DPReview photography website. We have seen that the sets of pro and con arguments between the arguments of the DSLR products and DPReview characterizations are similar in Table 7. Now we want to evaluate how good our experiential characterizations of products are, compared to those created by DPReview experts.

To do so, we define the function $\Phi(\overline{B}(p_i),\overline{B}(p_j))$, that estimates the degree in which a normalized bundle $\overline{B}(p_i)$ is better than another normalized bundle $\overline{B}(p_j)$:

$$\Phi(\overline{B}(p_i),\overline{B}(p_j)) = \frac{1}{2|C|}\sum_{k=1}^{|C|} s_k^i - s_k^j$$

where s_k^i and s_k^j are the sentiment values of arguments $\langle p_i, C_k, s_k^i\rangle$ and $\langle p_j, C_k, s_k^j\rangle$ in the normalized bundles of p_i and p_j. Φ is the average of these differences, a value in $[-1,1]$. If the value of $\Phi(\overline{B}(p_i),\overline{B}(p_j))$ is in $(0,1]$, it means that $\overline{B}(p_i)$ is better than $\overline{B}(p_j)$. If the value of $\Phi(\overline{B}(p_i),\overline{B}(p_j))$ is in $[-1,0)$, it means that $\overline{B}(p_i)$ is worse than $\overline{B}(p_j)$.

Using Φ, we create five rankings with the products of each camera corpus: one for each normalized bundle type $\overline{B}_G, \overline{B}_\sigma$ and $\overline{B}_F$, a DPReview ranking based on the DPReview overall score, and an Amazon ranking based on each product star rating. At the end, we have five different rankings for DSLR cameras ($\overline{B}_G$ rank, $\overline{B}_\sigma$ rank, $\overline{B}_F$ rank, DPReview rank and Amazon rank), five camera rankings for Compact cameras, and five camera rankings for Point & Shoot cameras.

In case two or more products of the same camera category had the same DPReview score, such as Olympus E620 and Nikon D3100, both cameras with a score of 72 out of 100, we only kept the product with most reviews, in this example the Nikon D3100. This left us with 10 different DSLR cameras, 10 Compact cameras, and 10 Point & Shoot cameras for the rankings.

Let us now compare the five DSLR rankings. The top 3 products for the $\overline{B}_G$ ranking are Nikon D7100, Pentax K-5 and SonySLT A-55. The top 3 products for $\overline{B}_\sigma$ are Nikon D7100, SonySLT A-99 and SonySLT A-55, and the top 3 ranked products for $\overline{B}_F$ are Nikon D7100, SonySLT A-99 and Pentax K-5. Notice that Nikon D7100 is the top product in the three bundle types. Nikon D7100 is also 1st (with a score of 85 points) in the DPReview ranking, followed by SonySLT A-99 and Pentax K-5. The top products of Compact and Point & Shoot camera categories are also similar for the 3 bundle types and DPReview.

Table 8 shows the Spearman rank correlation of the 3 rankings of bundle types with the DPReview score ranking and the Amazon star ranking, for DSLR, Compact and Point & Shoot cameras. We added a 6th random ranking to facilitate comparison. The random ranking correlation was obtained by averaging the Spearman correlations of 10.000 randomly generated product rankings with DPReview ranking and Amazon ranking. The 'Avg. DPReview Ranking' of Table 8 shows the average Spearman rank correlation between the DPReview ranking and the Amazon ranking of the three camera categories.

The results show that, for the DSLR camera category, the $\overline{B}_G$ ranking has the highest Spearman correlation with DPReview ranking (correlation of 0.80), followed by the cardinality ranking $\overline{B}_F$ (correlation with

	DPReview Rank	Amazon Rank
DSLR Spearman Rank Correlation		
$\overline{B}_G$ Ranking	**0.80**	0.39
$\overline{B}_\sigma$ Ranking	0.70	0.62
$\overline{B}_F$ Ranking	0.76	0.61
Compact Spearman Rank Correlation		
$\overline{B}_G$ Ranking	0.65	0.54
$\overline{B}_\sigma$ Ranking	0.62	0.55
$\overline{B}_F$ Ranking	**0.68**	0.57
Point & Shoot Spearman Rank Correlation		
$\overline{B}_G$ Ranking	0.57	0.33
$\overline{B}_\sigma$ Ranking	**0.86**	0.32
$\overline{B}_F$ Ranking	0.83	0.16
Avg. Random Ranking	0.27	0.27
Avg. DPReview Ranking	1	0.23

Table 8

Spearman rank correlation between the top 10 DSLR, Compact and Point & Shoot bundle rankings with DPReview product ranking and Amazon star ratings ranking.

DPReview of 0.76). For Compact cameras, $\overline{B}_F$ ranking obtained the highest Spearman correlation with DPReview ranking (correlation of 0.68). Finally, for Point & Shoot cameras, the bundle ranking that obtained the best correlation with DPReview ranking was $\overline{B}_\sigma$ (correlation of 0.86), closely followed by $\overline{B}_F$ (correlation of 0.83). These values tell us that there is a very strong correlation between DPReview and the rankings of our three bundle types, a good indicator of the high quality of the bundles. This is specially true with the rankings created from the cardinality bundles $\overline{B}_F$, which obtained an average Spearman correlation of 0.76 between the DSLR, Compact and Point & Shoot cameras and the corresponding DPReview rankings. The correlations for $\overline{B}_\sigma$ and $\overline{B}_G$ rankings are also strong, being notably higher than the random ranking correlations.

On the other hand, notice that the ranking correlations between the bundle rankings and Amazon star rankings are in comparison lower than the average Spearman correlation between the bundle rankings and the DPReview ranking. In fact, the average Spearman correlation between the three bundle rankings and the Amazon ranking is around 0.40, showing there is no strong similarity between the star-rating ranking and the rankings of the bundles acquired from the reviews. Furthermore, the Amazon star-based ranking does not correlate with the DPReview score ranking either (Spearman correlation of 0.23), indicating that there exists a notable difference between the star rating ranking of Amazon and the DPReview ranking. In fact, the Amazon ranking has a higher average correlation with the random ranking (0.27) than with Point & Shoot $\overline{B}_F$ ranking (0.16). These results seem to indicate that two people expressing similar arguments about a product can give different star-rating values (as an overall score). Nevertheless, the fact is that Amazon's star rating seems unsuitable to test the quality of the argument bundles.

6. Related Work

There are numerous applications that gather knowledge from user reviews, usually oriented to help other users make more informed decisions in the area of recommendation systems and CBR. The most common approach consists in characterizing a set of products by considering product aspects (also called features) mentioned in the reviews [1,2]. In this process, the set of aspects selected to characterize a product together with the sentiment analysis of the sentences have a crucial role in the final recommendation [13,3,6]. A related work on creating BLC is [7], but they manually tag a corpus with the classes (concepts) to which words belong, and then use supervised learning —while we discover the BLCs in an unsupervised way.

Another focus is identifying the sets of aspects with higher positive/negative polarity to give insights into the reason why items have been chosen [9]. Those approaches need previously to group the aspects to reduce the granularity in order to provide useful recommendations, often solved by clustering aspects using background knowledge to simplify the process. Our approach is different in the sense that we create basic level concepts [11] by exploring the usage of the aspects among the user reviews in an unsupervised way.

Using these basic level concepts, we build the bundles of arguments by identifying the pro and con concepts over the set of reviews of a product. Finally, we define a concept-wise satisfaction degree of a user query over a set of bundles using the notion of fuzzy implication [5]. This bundles both characterize the main pros and cons and supports reusing experiences of other people to the needs a specific user.

7. Conclusions and Future Work

In this paper we extend our previous work on aspect extraction and sentiment analysis and propose an unsupervised method to create a vocabulary of basic level concepts with the appropriate granularity to characterize a set of cameras. This concept vocabulary is useful to practically reuse other people's experiences with digital cameras because it abstracts the concrete terms used in the reviews as given by the aspect extraction approach. By analyzing the usage of the concepts over the reviews of each product, we find those concept occurrences that have a clearly positive or negative polarity and create their argument bundles. We present three different types of argument bundles, based on different aggregation criteria, each one defining the pros and cons of a product. The argument bundles allow us to define a satisfaction degree, interpreted in fuzzy logic and modeled with a fuzzy implication operator, between products and a user query.

An evaluation of the three types of argument bundles is performed and compared with the expert descriptions of the DPReview website, showing that the bundles of arguments identify pros and cons very similar to those used in DPReview. Moreover, the cardinality bundle ranking proved to correlate with the overall DPReview score ranking over the subset of the most frequent products, while Amazon star rating ranking does not correlate with either of them.

The characterization of products by means of the bundles of arguments and BLC is promising. We have observed that the quality of a product bundle is related to the quantity of reviews of that product: the products with more reviews have a richer vocabulary of pro and con arguments, while products with fewer reviews had more moots. This can be due to two reasons that open new lines for future work. First, improving the detection of aspects (for instance, considering also 3-gram aspects) could improve the argument bundles of those products with fewer reviews. Second, improving the sentiment analysis of reviews by developing a domain specific sentiment dictionary for digital cameras could enhance the accuracy of the arguments' sentiment.

Acknowledgements

This research has been partially supported by NASAID (CSIC Intramural 201550E022). We thank Lluís Godo and Pere García for their insightful comments.

References

[1] L. Chen, G. Chen, and F. Wang. Recommender systems based on user reviews: the state of the art. *User Modeling and User-Adapted Interaction*, 25(2):99–154, 2015.

[2] R. Dong, M. Schaal, M. O'Mahony, K. McCarthy, and B. Smyth. Mining features and sentiment from review experiences. In *Case-Based Reasoning Research and Development*, pages 59–73, 2013.

[3] R. Dong, M. Schaal, M. O'Mahony, K. McCarthy, and B. Smyth. Opinionated product recommendation. In *Case-Based Reasoning Research and Development*, volume 7969 of *Lecture Notes in Computer Science*, pages 44–58. Springer, 2013.

[4] X. Ferrer. *Concept Discovery and Argument Bundles in the Web of Experiences*. PhD thesis, Universitat Autonoma de Barcelona, 2017.

[5] P. Hájek, L. Godo, and F. Esteva. A complete many-valued logic with product-conjunction. *Archive for mathematical logic*, 35(3):191–208, 1996.

[6] M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proc. of the National Conf. on Artificial Intelligence*, pages 755–760, 2004.

[7] R. Izquierdo, A. Suárez, and G. Rigau. Word vs. class-based word sense disambiguation. *J. Artif. Intell. Res. (JAIR)*, 54:83–122, 2015.

[8] L. Meng, R. Huang, and J. Gu. A review of semantic similarity measures in wordnet. *Int. Journal of Hybrid Information Technology*, 6(1):1–12, 2013.

[9] K. Muhammad, A. Lawlor, R. Rafter, and B. Smyth. Great explanations: Opinionated explanations for recommendations. In *Case-Based Reasoning Research and Development*, pages 244–258. Springer, 2015.

[10] E. Plaza. Semantics and experience in the future web. In *Advances in Case-Based Reasoning: 9th European Conference, ECCBR 2008*, volume 5239 of *Lecture Notes in Artificial Intelligence*, pages 44–58. Springer Verlag, 2008.

[11] E. Rosch. Human categorization. *Studies in cross-cultural psychology*, 1:1–49, 1977.

[12] S. Sani, N. Wiratunga, S. Massie, and R. Lothian. Term similarity and weighting framework for text representation. In *Case-Based Reasoning Research and Development*, pages 304–318. Springer, 2011.

[13] P. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. Annual Meeting on Association for Computational Linguistics*, pages 417–424, 2002.

[14] W. Winkler. The state of record linkage and current research problems. In *Statistical Research Division, US Census Bureau*. Citeseer, 1999.

[15] Y. Y.Chen, X. Ferrer, N. Wiratunga, and E. Plaza. Sentiment and preference guided social recommendation. In *Case-Based Reasoning Research and Development*, pages 79–94. Springer, 2014.

[16] Y. Y.Chen, X. Ferrer, N. Wiratunga, and E. Plaza. Aspect selection for social recommender systems. In *Case-Based Reasoning Research and Development*, pages 60–72. Springer, 2015.

[17] S. Yitzhaki. Relative deprivation and the Gini coefficient. *The Quarterly Journal of Economics*, pages 321–324, 1979.