Making Sense of Ratings: A Common Quantitative Feedback Ontology

Florian Marienfeld
Fraunhofer FOKUS, Berlin
florian.marienfeld@fokus*

Edzard Höfig
Fraunhofer FOKUS, Berlin
edzard.hoefig@fokus*

Andrea Horch
IAT University of Stuttgart
andrea.horch@iat.uni-stuttgart.de

Maximilien Kintz
IAT University of Stuttgart
maximilien.kintz@iat.uni-stuttgart.de

Jan Finzen
Fraunhofer IAO, Stuttgart
jan.finzen@iao*

ABSTRACT

This paper proposes a common ontology for ratings, i.e. for quantitative user feedback data. Such a framework allows for semantic interoperability of data that adheres to it, which in turn enables the re-use, by making it independent from the original system.

In contrast to prior attempts to establish an unambiguous vocabulary, this approach introduces two components that are in our view necessary to formally understand what a user's rating actually means. The first is the aspect or facet, i.e. the viewing angle that was chosen to look at the rated thing. The second is the meta-model of scales following the scales of measurement that are widely used in descriptive statistics. So in plain words, we allow to formally specify how many out of how many score points something gets and with regards to what.

We follow the open world assumption of the Web Ontology Language (OWL) and design a vocabulary that is not specific to any domain. In turn, we rely on the premise that all domain specific concepts are available as semantic web resources with appropriate URIs.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Ontologies, Semantic Networks*

General Terms

Standardization, Measurement

Keywords

semantic web, ratings

*.fraunhofer.de

I-SEMANTICS 2011, 7th Int. Conf. on Semantic Systems, Sept. 7-9, 2011, Graz, Austria

1. INTRODUCTION

A great variety of applications make use of ratings, i.e. users explicitly stating their impression of something. These include recommendation systems, product search, service selection, information retrieval, and trust management to mention just the most prominent ones. Many researchers have explored the question of how to exploit rating data while fighting attacks/tampering. However, little work has been carried out to establish a common ontological framework for all rating systems. That is the focus of this article.

The key problem of this state of affairs is that all rating data are isolated in the very system where they are generated. That means a specific object may get rated on several platforms, but the scores can hardly be accumulated into one average, since all platform hosts have slightly or significantly different concepts regarding their feedback component. Thus, there is a need for a formal way to define what it means that a user picked a certain amount of stars after her purchase.

The remainder of this paper is organized as follows: Section 2 argues why such an ontology is needed. The actually proposed concepts are presented and discussed in Section 3. In Section 4 we argue that our proposal is suitable for real world domains by showing that popular rating systems can be modeled adhering to our proposal. We position our work among similar approaches in Section 5 and conclude in Section 6.

2. MOTIVATION

There are two key motivations for a common rating ontology: Interoperability and semantic technologies. In practically all major existing rating systems data is insulated from other platforms, e.g. a rating on a book at Amazon.com is no use to customers of any other book seller. This is, of course, due to economic reasons, but even several web sites wanted to cooperate, they would have to converge on common semantics in the first place. An example for this would different open knowledge platforms that allow users to rate items, like Wikipedia.org. Moreover, current rating data can only be used in the way it was designed. It is hard to apply them to fresh ideas or combine them in a new way, as there is no absolute meaning to most rating data.

A common ontology as presented here would remedy both problems. All rating data that adheres to it, i.e. that represents instances of the ontology's classes, could be used by semantic applications that span rating systems of different platforms.

The aim of this ontology is to make quantitative opinions of agents about things formally understandable. Thus, the guiding question in designing the concepts was "What information is necessary to be able to interpret a rating?" So in addition to the straightforward combination of a rater, a thing rated and a value, we decided it is necessary to define the aspect which was considered and the scale on which the value was assigned.

As a consequence of this guiding principle we did not model qualitative aspects such as free text answers. The purpose of free text or reviews is precisely to allow for those answers that a system is not designed to understand, hence they cannot be exploited in the semantic web in a statistical fashion, as is the case with quantitative rating. For a bridge from reviews to the semantic web see [Heath and Motta, 2008] and revyu [Heath and Motta, 2008].

3. ONTOLOGY

The ontology was modeled on top of the Web Ontology Language (OWL) [McGuinness and van Harmelen, 2009] and is reproduced here in Manchester syntax [Horridge and Patel-Schneider, 2009] for clarity. Its visualization in Figure 1 is helpful, but does not reflect all details defined in the code.

Central concept is the class Rating:

```
Class: Rating
    SubClassOf:
        about exactly 1 owl:Thing,
        submittedBy exactly 1 Agent,
        creationTime exactly 1 dateTime
        hasAspect exactly 1 Aspect,
        hasScale exactly 1 Scale,
        hasValue exactly 1 Value,
        (hasScale some NominalScale)
        or (hasValue some NumericalValue)
```

In OWL a subclass is a plain subset of its superclass. So if a class has multiple superclasses it is a subset of their intersection. In this case the superclasses are anonymous. The definition can be read as follows: All instances of the class, i.e. rating objects, must be in an "about" relation with exactly one Thing, Thing being the superclass of all other classes. Likewise, each instance must have exactly one submitting Agent, one time of creation, one Aspect, one Scale and one Value. The last two lines denote a union of the objects that have an OrdinalScale and those that have a NumericalValue. This constraint makes sure that rating objects cannot have a scale that requires a numerical value but not actually have a NumericalValue.

For completeness, the respective properties are listed here, along with their domains and ranges:

```
ObjectProperty: about
    Domain: Rating
    Range: owl:Thing

ObjectProperty: submittedBy
    Domain: Rating
    Range: Agent

ObjectProperty: hasAspect
    Domain: Rating
    Range: Aspect

ObjectProperty: hasScale
    Domain: Rating
    Range: Scale

ObjectProperty: hasValue
    Domain: Rating or Scale
    Range: Value

DataProperty: creationTime
    Domain: Rating
    Range: dateTime
```

No further constraints are imposed on the class Agent. Anything that is typed as such can be submitter of a rating. In most cases an instance of Friend-of-a-friend's[1] class Agent will be appropriate, but it is no use at this point to enforce this, as other suitable agents can be thought of.

```
Class: Agent
```

The concept Aspect is used to declare with respect to what the rating was issued. Combined with the recursive link isRefinementOf, this allows for a hierarchically refined structure or taxonomy of aspects. Simple examples are "speed of delivery" or "friendliness of staff of support". This requires that the referenced resources have valid URIs. Suitable repositories for that can be Linked Open Data collections like DBPedia [Auer et al., 2007] or makeshift approaches like disambiguated Wikipedia pages as suggested in [Milicic, 2008].

```
Class: Aspect
    SubClassOf:
        isRefinementOf max 1 Aspect

ObjectProperty: isRefinementOf
    Domain: Aspect
    Range: Aspect
```

The design of the scales is based on the "scales of measurements" that are elementary in statistics [Stevens, 1946]. A comparison is presented in Table 1. With the help of the appropriate scale it is possible to state how one rating value can be compared and aggregated with other values. To simplify property definitions scales are differentiated into discrete and continuous:

```
Class: Scale
    EquivalentTo: ContinuousScale or DiscreteScale
```

Let us consider discrete scales first:

[1] FOAF: http://xmlns.com/foaf/spec

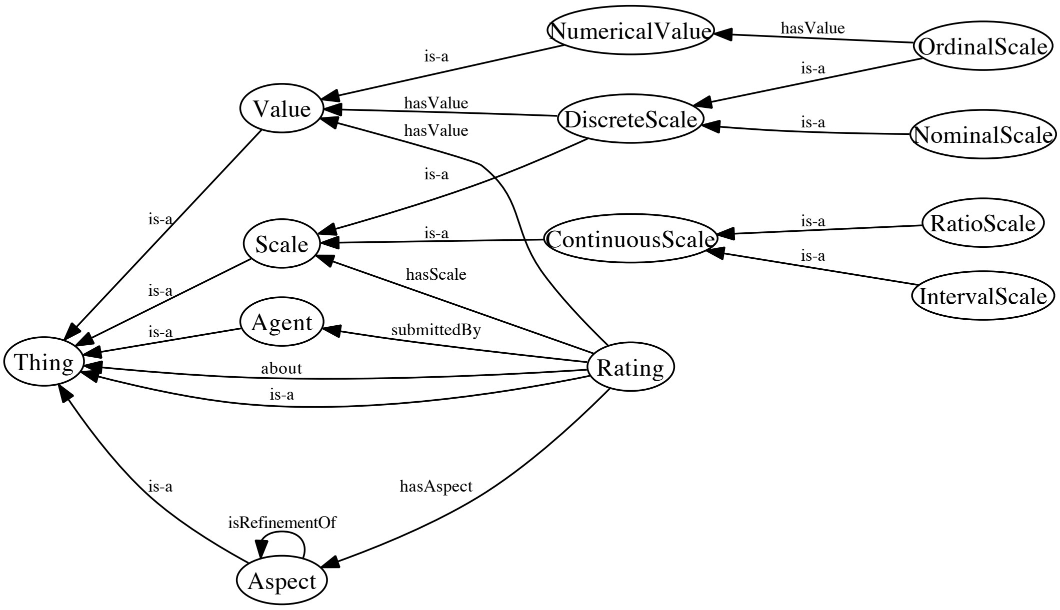

Figure 1: Visualization of the Rating Ontology

```
Class: DiscreteScale
    EquivalentTo: NominalScale or OrdinalScale
    SubClassOf:   Scale
    DisjointWith: ContinuousScale

ObjectProperty: hasValue
    Domain: Rating or Scale
    Range:  Value

Class: Value
```

They have a fixed set of resources to choose from when rating a thing. Examples for values are URIs for “red”, “very much”, “yes”, or “negative”. The two legal subclasses NominalScale and OrdinalScale indicate whether or not the values are unambiguously comparable. It may be that the NominalScale is rarely used, but it was included nevertheless for completeness to cover cases like “favorite vehicle for a route - bike, car, walk“ and situation where only one option is available (like “like”). The order on an OrdinalScale is ensured by the exclusively numerical values associated with it. Despite the assigned number, no meaning can be derived from it other than the ordering. An example of this is “accuracy of description - inaccurate (0), accurate (1), very accurate (2)”. Clearly, inaccurate is less than both the other, but it does not follow that the distance from inaccurate to accurate is as big as from accurate to very accurate. This difference intuitively leads to the corresponding legal computations indicated in Table 1. Nominal values only allow for computation of the most frequent element (mode), ordinal values also imply a “middle” element (median).

```
Class: NominalScale
    SubClassOf:
        DiscreteScale
        hasValue min 1 Value
    DisjointWith: OrdinalScale

Class: OrdinalScale
    SubClassOf:
        DiscreteScale
        hasValue only NumericalValue
        hasValue min 2 Value
    DisjointWith: NominalScale

Class: NumericalValue
    SubClassOf:
        Value,
        hasFloat exactly 1 float
```

In contrast, a ContinuousScale can accommodate any value in a certain range. Knowing what range a value was chosen from is essential for interpreting the meaning of a rating. Obviously, the values have a natural order and, in addition, distances have a meaning. The crucial difference between an IntervalScale and a RatioScale is that on the former ratios are not applicable, but they are on the latter. In other words, on an IntervalScale only the distance between values is of meaning, where as on RatioScales also the distance from zero has a meaning, i.e. the absolute value. This difference is also relevant for the semantics of a feedback value. An example of a RatioScale would be “hours of play time” for a computer game. In this case a value has an absolute meaning, independent of the range and an expression like “20% more play time” do make sense.

Scale		applicable expressions	applicable average
Discontinuous	Nominal	-	mode
	Ordinal	$a > b$	median
Continuous	Interval	$a > b,\ a - b$	arithmetic mean
	Ratio	$a > b,\ a - b,\ a/b$	geometric mean

Table 1: Scales of Measurement

```
Class: ContinuousScale
    EquivalentTo:
        IntervalScale
        or RatioScale
    SubClassOf:
        Scale,
        lowerBound exactly 1 float ,
        upperBound exactly 1 float
    DisjointWith: DiscreteScale

DataProperty: upperBound
    Domain: ContinuousScale
    Range: float

DataProperty: lowerBound
    Domain: ContinuousScale
    Range:  float

Class: IntervalScale
    SubClassOf: ContinuousScale
    DisjointWith: RatioScale

Class: RatioScale
    SubClassOf: ContinuousScale
    DisjointWith: IntervalScale
```

Some may feel the urge to define at some point what is a good and what a bad rating. However, good or bad can in some cases only be judged by the interpreter of a rating (consider e.g. the aspect "complexity" – a high complexity may be great for thriller movie enthusiasts). Therefore, we follow the following intuition: the higher the value, that more applicable is the respective aspect.

4. SUITABILITY FOR USE IN PRACTICE

There are several websites dealing with ratings. One popular example is Amazon, where consumers have the possibility to rate the books or other products they have bought and sellers rate buyers.

This section will show how things can be rated on different websites, which rating systems and which scales are used, in order to argue that all these rating systems can be modeled using our proposed rating ontology. To support that claim we looked for a broad variety of different rating systems. The rating systems on the considered websites were classified and categorized to a scale of measurement. The results of this analysis are summarized in Table 2.

In the course of that, we grouped the analyzed rating systems into the following six main rating system types, which are used on the websites as specialized subtypes (e.g. main type: star system ⇒ subtype: 6 star system). These groups follow the categorization in [Sparling and Sen, 2010]. We also identified aspects where the object is rated specifically by several aspects like "kindliness" for a craftsman.

1. *Unary System*
   In unary rating systems a rater can only choose one value for doing the rating. In most cases unary rating systems are positive rating systems where a user or customer can express if he likes a thing (e.g. Facebook's "Like it"-Button).

   The scale of unary systems is a nominal scale, because the user can just express if he agrees with a given value (if he does not agree he will do nothing). The rating only counts the votes of agreement with the given rating value, it does not say anything about the votes of disagreement.

2. *Binary System*
   In a binary rating system the rater can express if he likes a thing and he also can express if he does not like it. There are two rating values for doing the rating. Usually there is a value a rater can choose to express a linking and one to express a disliking (e.g. Youtube "Top/Flop"-Buttons) vote.

   These characteristics meet the nature of an ordinal scale, because one can sequence the votes, but one can not say anything about the distance from a good to a bad rating value. A sum or percental rate of the ratings for each value can be built (how many positive and how many negative ratings are given to a thing).

3. *Sentiment System*
   The sentiment rating system delivers a more extended view on the rater's opinion. Usually it offers three values like "negative", "neutral" and "positive"(e.g. eBay rating of sellers). This kind of rating system offers an easy and fast overview of the sum or percental rate of "positive", "neutral" and "negative" ratings given to a thing.

   Sentiment rating systems like binary rating systems also use an ordinal scale to build a sum or percental rate of the ratings for each given value.

4. *Grade Point System*
   A grade point rating system is inspired by the grade point system of schools. In many countries letters A, B, C, ... are used to label grades, where A is usually the best. In Germany numbers from 1 to 6 are used instead, where 1 is the best grade. Some websites like spickmich.de – where teachers and schools are rated – are using this kind of rating system.

   Grade point systems are using an ordinal rating scale, which can be different between cultural groups, because they are using different school grade point systems. To fit these system in our ontology one could argue that some numeric value with meaning must be assigned to each label. In the German case even the grade numbers must be considered labels requiring a mapping to semantic values. As a matter of fact,

Platform	*Rated Object*	*System type*	*Denotation*	*Scales*	*Range*	*Aspects*
Holidaycheck.de	Hotels	Star System	6 Star System	Interval	1 - 6 suns	hotel, room, service, location, gastronomy, sport
		Slider System	Recommendation rate	Interval	0 - 100%	
YouTube.com	Online-Videos	Binary System	Top-Flop System	Ordinal	Flop-Top	
Amazon.com	products	Star System	5 Star System	Interval	1 - 5 stars	
	Sellers	Star System	5 Star System	Interval	1 - 5 stars	
eBay.com	Sellers	Sentiment System	Sentiment System	Ordinal	positive/ neutral/ negative	accuracy of description, satisfaction with communication, delivery speed, delivery costs
	Consumers	Unary System	Positive System	Nominal	positive	
Facebook.com	Articles/ Facebook-Sites	Unary System	Positive System	Nominal	Like it	
maps.google.com	Restaurants/ Companies/ etc.	Star System	5 Star System	Interval	1 - 5 stars	
MyHammer.de Branchenbuch	Craftsmen/ Orderer	Star System	5 Star System	Interval	1 - 5 stars	reliability, kindliness, craftsmen: quality, buyers: payment behaviour
	Craftsmen/ Buyer	Sentiment System	Semtiment System	Ordinal	positive/ neutral/ negative	
spickmich.de	Teachers/ Teacher's citations/ Schools	Grade Point System	Grades	Ordinal	Grades 1 - 6	schools: teacher's quality, buildings, technical equipment, atmosphere, range of subjects/courses, administration, sport offers, cancellations of classes, dining options, participating in decision taking

Table 2: Examples of rating systems

an official mapping exists, that maps grade "1" to 15 score points and grade "6" (fail) to 0 score points, which is well in line with the semantics of high and low described in the previous section.

5. *Star System*
   Star rating systems are the most often used rating systems. The count of stars for the rating differs between the different websites, but the 5 star rating system seems to be the most popular. Some examples are Amazon, Holidaycheck, MyHammer and Google maps (see Table 2).

   Star rating systems are using an interval scale, because one can accumulate the ratings as well as build an arithmetic mean; additionally, the distance from one value of the scale to another are meaningful.

6. *Slider System*
   A slider rating system offers the most possible values to rate a thing or a special aspect of a thing. Sometimes you can choose values between 0 and 100. One popular website using this rating system is Holidaycheck were people can rate their hotels after they have stayed there for there holidays.

   Like the star rating system the slider system is also using an interval scale to be able to accumulate the ratings and to build arithmetic means.

The above presented rating systems differ in the count of possible values. That means a unary system allows only one possible value for the rating, a binary allows two values, a grade point system in Germany allows six values, a star system allows zero to less than ten values and a slider system usually allows zero to 100 values. It is also possible for each system to rate a thing by rating several aspects of the thing by using the scale of the system.

We have identified six different types of rating systems which are using one out of three different scales of measurement. Most of them have a minimum value and a maximum value for defining the best and the worst possible rating. None of the systems under investigating presented a concept that is not covered in our proposed ontology.

5. RELATED WORK

There are several publications that have rating interoperability as a theme, but to our knowledge none brings all insights to the semantic web. We considered strength and weaknesses of the these to optimize our proposal.

In tvblob [Longo and Sciuto, 2007] the authors present an ontology for rating. However it can not describe under what aspect a thing has been rated and what scale was used. Possibly due to this, tvblob has failed to attract significant attention.

The Review Vocabulary[2] and revyu [Heath and Motta, 2008] are more elaborate but focused rather on reviews, i.e. qualitative feedback. Quantitative ratings are not modeled with exhaustive detail.

Another similar ontology is presented in [Sriharee, 2006]. It is however geared towards ratings by agencies. Hence, it is not intended/suitable to turn user feedback into machine understandable facts. Rather, it allows service providers on a technical level to incorporate third party classifications in their profiles.

In [Grinshpoun et al., 2009] cross-domain interoperability of reputation is discussed but no ontology is presented.

Before the semantic web even became popular a rating language was formulated[3]. However, only the syntax was considered, there is no way to derive absolute meaning.

6. CONCLUSION

We have proposed a common ontology for quantitative ratings, which is in our view complete, yet universal. Operators who are aware of it may model their rating aspects, scales and values accordingly. Once the semantics are settled in this way, they may design their feedback forms accordingly. The resulting user rating can, thus, be understood by both humans and machines and used in unforeseen fashions. Obviously, this requires that the ratings are available as open data.

This may not be the final version of the ontology. Rather this paper can be seen as a Request for Comments of the semantic web. That means, with the help of interested ontologists a refined and consolidated version may later be published for a wider adoption.

This work inspires two main lines of further research, which is intended by the authors. One is the relation between incompatible rating components, i.e. the questions whether transformation from one scale to another is sensible and how subaspects sum up to more general aspects. The second line of thought can be phrased as access control, i.e. how to specify who is allowed to give/read a rating and privacy issues, including anonymity or pseudonymity of raters.

7. ACKNOWLEDGMENTS

This work was partially funded by the German THESEUS/TEXO programme and by the openXchange project of the German Federal Ministry of Economy and Technology under the promotional reference 01MQ09011. It was inspired by the TEXO service ontology rating module by Tom Kiemes and Daniel Oberle of SAP Research, Karlsruhe.

8. REFERENCES

[Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., and Cudré-Mauroux, P., editors, *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer Berlin / Heidelberg.

[Grinshpoun et al., 2009] Grinshpoun, T., Gal-Oz, N., Meisels, A., and Gudes, E. (2009). CCR: A Model for Sharing Reputation Knowledge Across Virtual Communities. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '09, pages 34–41, Washington, DC, USA. IEEE Computer Society.

[2] `http://vocab.org/review`

[3] `http://www.w3.org/TR/REC-PICS-services/`

[Heath and Motta, 2008] Heath, T. and Motta, E. (2008). Revyu: Linking reviews and ratings into the Web of Data. *Web Semantics: Science Services and Agents on the World Wide Web*, 6(4):266–273.

[Horridge and Patel-Schneider, 2009] Horridge, M. and Patel-Schneider, P. F. (2009). OWL 2 Web Ontology Language - Manchester Syntax. Technical report, W3C.

[Longo and Sciuto, 2007] Longo, C. and Sciuto, L. (2007). A Lightweight Ontology for Rating Assessments. In *4th Italian Semantic Web Workshop - Semantic Web Applications And Perspectives (SWAP)*, volume 134, Bari. Citeseer.

[McGuinness and van Harmelen, 2009] McGuinness, D. L. and van Harmelen, F. (2009). OWL Web Ontology Language. Technical report, W3C.

[Milicic, 2008] Milicic, V. (2008). Case Study: Semantic tags. Technical report, W3C.

[Sparling and Sen, 2010] Sparling, E. I. G. and Sen, S. (2010). Cognitive Load of Rating Scales. Honors Project.

[Sriharee, 2006] Sriharee, N. (2006). Semantic Web Services Discovery Using Ontology-Based Rating Model. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 608–616.

[Stevens, 1946] Stevens, S. S. (1946). On the Theory for Scales of Measurement. *Science*, 103(2684):677–680.