Egret Endeavors, Inc.
775 E. Blithedale Ave., #217
Mill Valley, CA 94941
1 (888) 347-3836
www.egretendeavors.com

ATTRIBUTE PROFILING:

The evolution of collaborative filtering

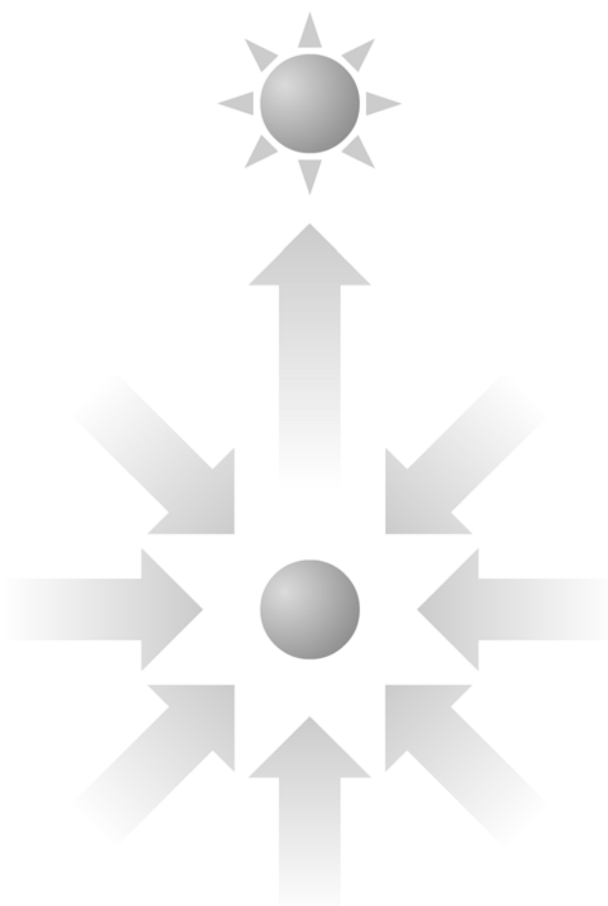


TABLE OF CONTENTS

OVERVIEW

As you know the ability to accurately recommend a product to a user is a huge thing. A small increase in accuracy can lead to substantial increases in revenue, not to mention happier customers. To increase recommendation accuracy is a text-book case of a win-win scenario.

Also, not only is it crucial to the user experience but, more importantly, it is a backbone service for many websites that can easily be the difference between long-term success or failure. It is not hard to see that a website that can accurately recommend products to its users will have a significant advantage over a website that can't. Would Google be anywhere it is today if it didn't start off with a search engine that was technically superior to its peers? Since you are reading this paper then I'm sure you feel the same way.

A website that can accurately recommend products to its users will have a significant advantage over a website that can't.

So, with that in mind, the purpose of this white paper is to show you how to dramatically increase your product recommendation accuracy.

However, this is not a proposal to merely increase your accuracy by a small percentage. This white paper is about a fundamental paradigm shift that is necessary to completely blow open the doors on the current Collaboration Filtering limits.

To that end this paper will take you through three steps:

- A brief Introduction
- The current problem
- Attribute Profiling: A high-level overview

Terminology

Before we begin, please note that I use the term "product" to refer to any type of item in which you can apply Attribute Profiling to (e.g. a DVD). A "product category", then, is a collection of the same type of products such as "movies", "CDs" or "books". Would you like to be put in contact with someone who shares the same interests and characteristics as you? It is interesting to note that "product" may refer to people, as well, in the case of social-networking or dating services.

For the sake of simplicity the primary product category of discussion will be movies.

A BRIEF INTRODUCTION

The collaborative filtering industry is in dire straits. Nothing is wrong, per se, but it has reached a point of critical mass in its current state.

Take Netflix, for example. Recently they have created a one million dollar public contest to whoever can increase their recommendation system by 10%. This public contest simply demonstrates that the collaborative filtering industry believes there is room for improvement but, at the same time, has no idea where that improvement is going to come from.

To all of this I'd like to propose a quote.

"Imagination is more important than knowledge." – Albert Einstein

I offer this quote simply because to see the fundamental barrier that collaborative filtering is against – and a solution around it – it is going to require a bit of imagination on your part. Not a lot, mind you, but a bit. A good dose of common sense will go a long way as well.

That being said let us begin. Like so many good movies, this white paper is based upon a true story…

A true story

Unless you know* why *someone likes (or dislikes) a product there is no way you can reliably recommend another to him/her.

A few years ago I was shopping on-line to purchase a new rock CD that I had recently heard. This particular CD has some positive, philosophical themes running through it that I particularly liked. It turned out to be a great lifesaver for when things got tough. The heavy rock guitar, combined with a positive message, was really powerful. If I ever needed a good kick in the rear then I would just put on this CD, turn up the volume, and blow away whatever funk my mood was in. In no time at all I'd have my positive, mental attribute back and I'd be ready to take on the world again.

However, in the days that followed, I began to notice the recommendations the website would give to me because of that purchase. What made me take notice of these recommendations, in particular, was how off they were! They offered lots of heavy rock bands that geared all of their content towards the stereotypical "sex", "drugs" and "rock 'n' roll". While I, personally, have no problem with bands that build their career around these themes it simply just wasn't what I was into at the time.

At that moment I had a small realization. It was small at the time but I had no idea it would lead to the entire foundation of this white paper. The realization was this:

Unless the on-line retailer knows *why* I bought my CD then there is no way they can reliably recommend another to me.

This concept is so important that I am going to say it again.

Unless you know *why* someone likes (or dislikes) a product there is no way you can reliably recommend another to him/her.

Yes, of course the industry has Keywords, Genres and Tags, but as we saw above with my rock CD, and as we'll see below, these are not enough.

THE CURRENT PROBLEM

"Houston, we have a problem."

To quote the New York Times article, *And if You Liked the Movie, a Netflix Contest May Reward You Handsomely*:

> "John Riedl, a professor of computer science at the University of Minnesota and a pioneer in the field of collaborative filtering, said that Netflix and Amazon now had the most advanced recommendation systems.
>
> 'Most of the easy stuff has been squeezed out already,' he said, adding that it had become increasingly difficult to make substantial progress in predicting accuracy.
>
> 'Any time you start working on any of these scientific or engineering problems, there's a period of dramatic improvement,' Professor Riedl said. 'It slows down because in a sense you're competing with 15 years of really smart people banging away at the problem.' "

Slows down? I'd say by the Netflix contest things have all but stopped. When the best companies in the world are throwing all of the resources behind collaborative filtering and they resort to a public contest (which, to Netflix's credit, is a brilliant idea given the current limitations) you know the industry is stuck.

Why is this? Let's take a look at the two leaders in the industry, Netflix and Amazon.

Plot Keywords, Genres and Tags (Oh my!)

Both companies have a user rate a product a five-star value and classify their products via genre. I'll refer to this as the "genre-rating data model" of collaborative filtering.

A genre comparison simply isn't enough to reliably recommend a movie.

Now, from the Netflix website, "The combination of genre ratings and movie ratings tell Netflix about your movie tastes and help us highlight the movies you're most likely to enjoy."

The idea is simple enough. Given two users, A and B, we compare user A's movie ratings to another user that has very similar ratings, user B. If user B happens to like movie X, and user A hasn't seen it yet, then we can recommend movie X to user A. Of course there are all kinds of proprietary algorithms within this concept to optimize the recommendations. However, like we saw above, if we don't know *why* user B liked movie X then how sure are we user A will like it?

Let's take the movie *Brazil* for a moment. If you don't already know, *Brazil* is a surreal romp through fantastic, sci-fi themes. Now, if user A liked *Star Wars* and *Star Trek* and user B liked *Star Wars*, *Star Trek* and *Brazil* how certain are we user A will like *Brazil*? Just because *Brazil* is classified as a Sci-Fi movie tells us very little about whether or not user A will like it. If you've ever seen *Brazil* then it's not hard to imagine that user A may not like it due to its very stylistic nature. In this case, in fact in *all* cases, a genre comparison simply isn't enough to reliably recommend a movie.

What do other sites do? Well the popular Internet Movie Database (IMDB) website uses Plot Keywords. However, those are not linked to a user, only the movie. And they're linked to the movie's plot at that, not necessarily *about* the movie (and can be quite random as well).

Amazon also uses user-defined Tags. While this is interesting, it allows any user to create any Tag about a product. The downside to this is there is no structure to link the web of random Tags together. Two users may love the same movie for the same reasons but if they use different Tags to describe the movie you can never derive that information.

You can't get there from here

The best method, so far, is the genre-rating data model and that falls short. There is not enough information within the genre-rating data model to achieve the desired results of a consistently successful product recommendation. It's like trying to spell the word "obsolete" using only consonants, no vowels – all you get is "bslt". No matter how hard you try, no matter how clever you get, you can never reach the goal. The given domain (consonants) cannot reach the desired range ("obsolete"). Likewise, the given domain of the genre-rating data model cannot reach the desired range of truly accurate product recommendations.

Until the way an industry recommends a product is in-synch with the way users relate to those products it will never get accurate recommendation results. How could it?

Attribute Profiling collects and organizes the necessary data to capture the "what" and "why" for products and users while, simultaneously, providing a seamless way to compare and match the two together.

What is needed

What is needed to overcome the fundamental limitation of the genre-rating data model is a way to capture *what* a product is like, *why* a user likes or dislikes a product and relate the two together. In short, new data must be collected. If a person likes or dislikes a genre then that is too general – there are, literally, thousands of products that can fit into that category. On the other hand, if a person likes, for example, Terry Gilliam movies staring Robert De Niro then that is too explicit – there is only one match (that I know of) that fits that criteria. What is needed is a balance between the two extremes – a list of simple, generic adjectives that span all products.

While a genre is a noun, and a star-rating is a value, what are needed to break the genre-rating data model are adjectives. Companies need to model their user/product data around adjectives to be able to truly offer accurate recommendations.

For the remainder of the paper instead of saying "adjectives" let's use another synonym: "Attribute".

Attribute Profiling, then, collects and organizes the necessary data to capture the "what" and "why" for products and users while, simultaneously, providing a seamless way to compare and match the two together.

ATTRIBUTE PROFILING: A HIGH-LEVEL OVERVIEW

What is it?

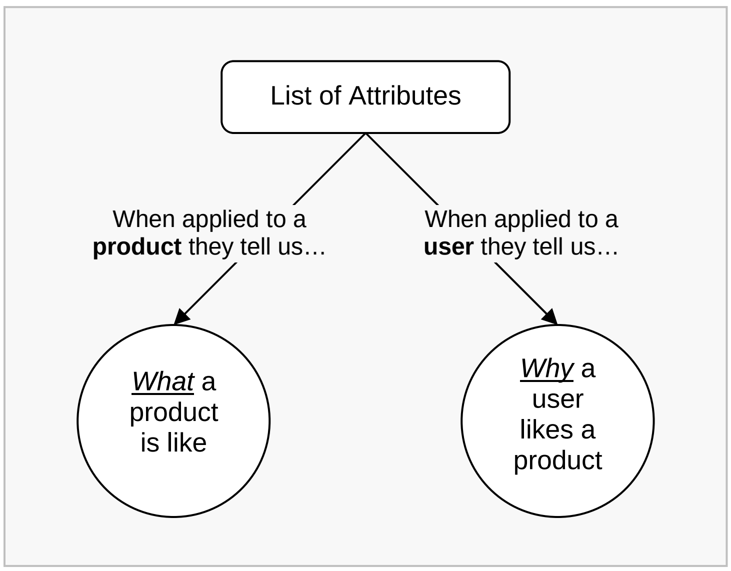
Attribute Profiling, in short, collects a pre-defined set of Attributes of a product, a user's preferences towards those Attributes, and allows a head-to-head comparison of both.

The "what" and "why" of Attributes

The Attributes applied to a product tells us *what* the product is like. They give us a weighted description of a product that goes beyond the black and white categorization of genres. For example, Attributes allow us to describe a movie as "full of action", "has a bit of romance", "a great ending" or "has mild gore". The values are not black and white, such as "has romance" or "does not have romance", but contain a weighted value relative to all of a product's Attributes.

Likewise, Attributes applied to a user tell us the *why* she likes or dislikes a product – it tells us her tastes, or "moods", of a particular product. For example, collecting user Attributes on movies is like having a user tell us she, "really likes action/adventure stories that are full of intrigue. My favorite, however, is a deep, romantic movie that has a happy ending – that is the best!" See Figure 1 below.

Figure 1



A profile of Attributes

Once we've defined our Attributes we can then begin to collect values for them relative to a product and/or user. From that data we can create a profile – a profile of Attributes.

Since the Attributes on a product and the user are the same set of Attributes we can easily compare them side-by-side to find the best match. We'll see how this works next.

How it works

Defining your Attributes

The first thing that must be done is to define a set of Attributes for a given product. For example, for movie and music products the Attributes could be:

Movie product Attributes:	Music product Attributes:
Action / Adventure	Experimental
Animation / Computer Graphics	Hard
Artistic / Surreal theme(s)	Jazz
Competition / Sports	Lounge
Drama	Melancholy
Gore	Pop
Humor	Psychedelic
Innocence / Youthfulness	Rap
Insights / Education	Rock
Intrigue / Suspense	Soft
Music	Soothing
Philosophy / Spirituality	Sophisticated
Poignancy / Feeling	Strong beat
Romance / Relationships	Trance
Sci-Fi / Fantasy theme(s)	Up-beat
Sexuality	
Story / Plot	
Violence	

You may notice that for both product categories (movies and music) that some Attributes are genres as well. The key difference to note is that these "genres", in the context of Attributes, are adjectives and not nouns. Attributes are concerned about adjectives and strive to describe a product by its various flavors, no matter how many there are. You could say that while genres are designed to categorize, and thus "lock down", Attributes capture a product's essence in a completely open-ended manner.

For example, in the music Attributes what genre does "soothing", "up-beat" or "sophisticated" belong to? Any and all genres and that is exactly what we want – adjectives, not nouns.

Figure 2

Collecting your Attributes

Once your Attributes are defined the next step is to collect them. Not to collect the Attributes, themselves, but the values of each Attribute as they relate to the products (the "what") and users (the "why").

The simple and most obvious way is to ask. At the same moment when a user would normally give a star-rating to a product you could simply have another section that asks the user to check off each Attribute of the product that they either liked or disliked. An Attribute rating GUI could look something like Figure 2.

The user is allowed to check off any Attributes that apply. When a user does this, on a per product basis, this is called a Profile submission. So what, exactly, does this do for us?

For each Profile submission (the Attributes a user applied to a product) we get two things: the "what" *and* the "why".

The "What"

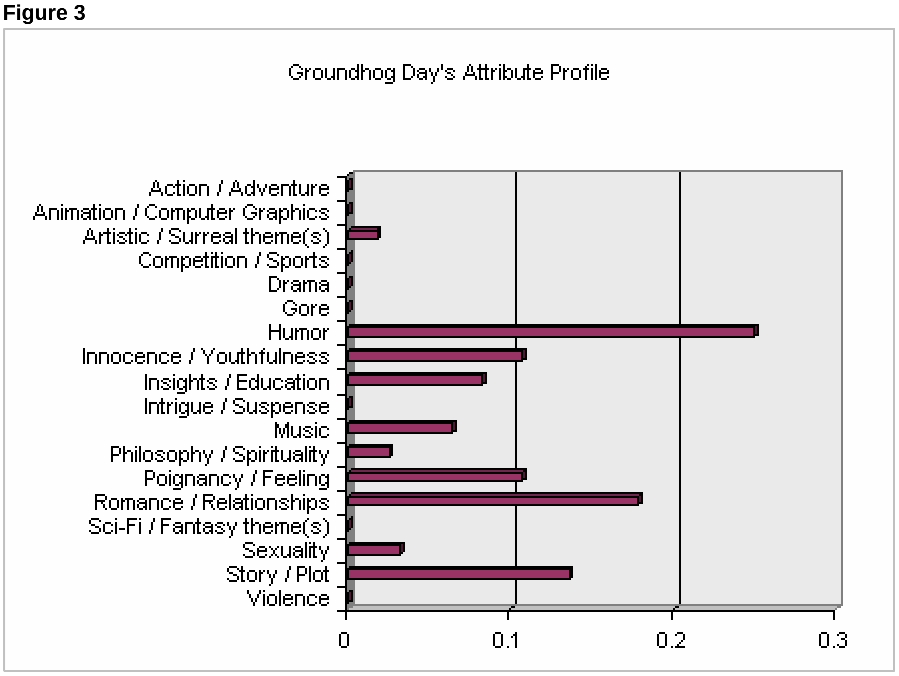
We get the "what" because if the user checked that he liked the Action/Adventure Attribute on a movie then that movie must have some positive action and/or adventure aspects to it. In that case we would add one to that movie's Action/Adventure Attribute and keep track of the running total. If he disliked some Attributes then those Attributes would have one subtracted from that Attribute's total. If some movies have really bad Attributes then those Attributes could end up with negative values.

Then, if you take the cumulative values of all the Attribute submissions of thousands of users, as a whole, you get the Collective's opinion. The Collective, then, defines what a movie (or product) is like – what Attributes are good, which are bad, etc… This is perfect because it doesn't matter whether or not the Collective's opinion is correct (if it were possible to determine so). Since they (the Collective) are our target audience it's the only opinion we care about.

Because each product may have a varying number of Attribute submissions we normalize the Attribute values to bring them all to the same scale. In essence, each Attribute takes the value of its relative percentage to the sum total of all the Attributes. Once this is done it becomes that product's Attribute Profile.

Figure 3 shows what a possible Attribute Profile could look like for the movie *Groundhog Day*.

Figure 3



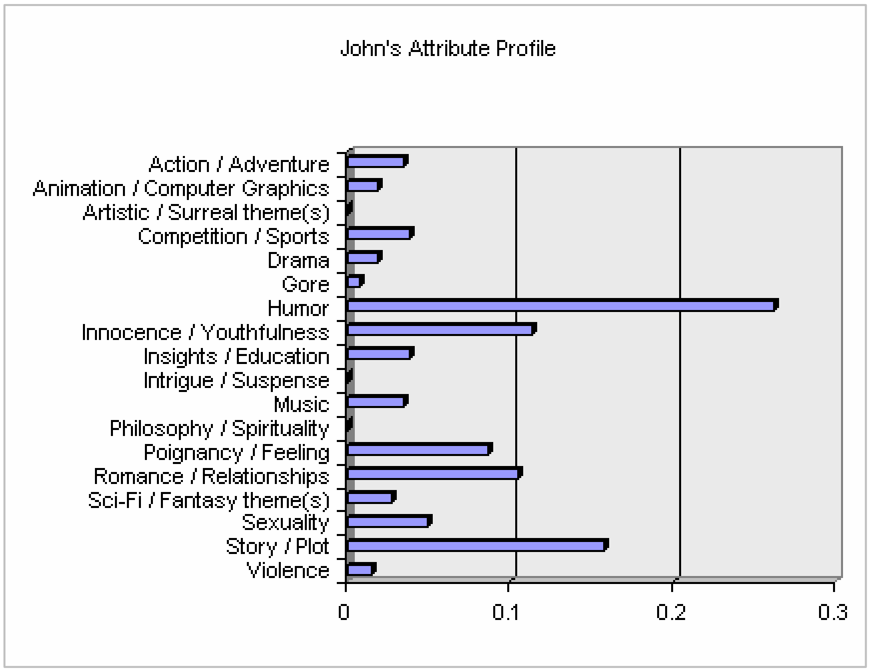
Notice how each Attribute has a different value. An Attribute Profile gives a complete and unique description of *Groundhog Day* – very much acting like its "fingerprint" – which then can be used to compare against a user's Attribute Profile.

The "Why"

The user's Attribute Profile, which gives us the "why", is derived from the exact same data that creates a product's Attribute Profile. When a user checks he likes a movie's Action/Adventure Attribute then, of course, he must like action and/or adventure. Similar to a product's Attribute Profile we add one to the user's Action/Adventure Attribute.

We keep a running total of all the Attributes a user submits over time and once we normalize the values we have that user's Attribute Profile. See Figure 4.

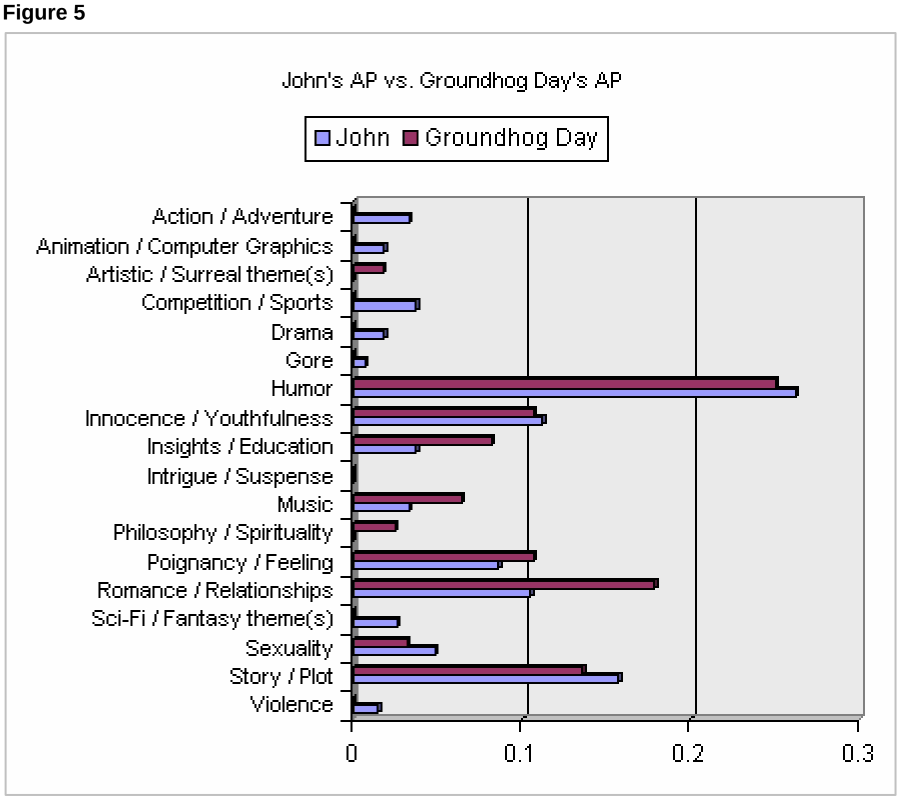
Figure 4



The Comparison

The final step is to compare a product's Attribute Profile to a user's Attribute Profile. The Attribute Profiles, normalized to the same scale, are then compared, like two "fingerprints", and you simply return the closest match out of a list of potential products.

Figure 5



From Figure 5 we can see that there is a strong possibility that John would like the movie *Groundhog Day*.

The obvious concerns

When looking at Attribute Profiling some immediate concerns come to mind:

- How do we Profile the thousands of existing products?
- Why do you believe users will submit Attributes for a given product?
- How does Attribute Profiling fit into our current collaborative filtering system?

Let's discuss each one at a time.

If you implement it the data will come

Before the ubiquitous five-star rating there were thousands of products that have not been reviewed. The need to offer product recommendations, along with the internet, gave birth to user ratings and, before long, every historical product had a collective rating of some sort.

The current need to offer superior product recommendations has given rise to the awareness that a new data paradigm is needed. Like the five-star ratings before it, simply by implementing Attribute Profiling for products will populate Attribute Profiles for all historical products over time.

The process is efficient as well because in one, single step each user is simultaneously contributing to the global, Attribute Profiling process for all products as well as building their own, personal Attribute Profile.

Success is the greatest attraction

It is natural to think that fewer users will take the time to submit Attributes for a given product than those who just submit a five-star rating. Initially this will most certainly be the case. However, not everyone submits a five-star rating and there is no indication that the industry lacks in those. Take IMDB for example, a site dedicated simply to the rating of movies, has millions of rating values and the users who submitted them gain absolutely nothing from them. They do gain, however, the satisfaction that their voice, no matter how small, has been heard. The desire to have one's opinion heard is very hard to resist, especially if there is an easy outlet for your opinion. To that end there are still thousands upon thousands of users who take the time and energy to actually *write out a review* of a given product. Again, these people have no hope that their review will ever be read or even considered amidst the flood of reviews, nor will they ever receive *any* benefit from their submissions, but they are happy to do it.

Is it so hard to imagine the majority of users taking an extra few moments to check off a few boxes to reflect their feelings on a product? Especially when it is so easy to do, right there below the familiar five-star rating?

Would Google be where it is today if it didn't, first and foremost, provide a better service?

If you make public a product's Attribute Profile, and the total count of submissions, then each user can actually see how their Attribute submissions is shaping the Profile of a product. Very similar to how an online vote increases by one when you cast your vote. However small, there is satisfaction of knowing, and seeing, how you have contributed to an end result.

In addition, you can show each user their Attribute Profile as well. Then each user can see, over time, how their own Profile changes as the more submissions they make and how it compares to their favorite products.

Finally, if all of that eye-candy is not enough, there is no greater attraction for a user than knowing the recommendations he/she gets are, truly, *really good* recommendations. The more submissions a user makes the more accurate the product recommendations become and the better service that user receives.

Again, would Google be where it is today if it didn't, first and foremost, provide a better service?

Extension, not replacement

As we can see from our examples above Attribute Profiling is not a complete replacement of any existing collaborative filtering techniques such as the genre-rating data models. If genre is a well established means of organizing your product data then Attribute Profiling can simply rest inside of that structure. In fact, especially in the case of genre, those structures can be very beneficial as described in the corollary white paper *Attribute Profiling: A complete walk-through*.

Attribute Profiling allows for a much deeper and richer data analysis and comparison.

The same thing can be said of Keywords or Tags. Whatever mechanism is currently in place, Attribute Profiling allows for a much deeper and richer data analysis and comparison within that mechanism. In short, Attribute Profiling will extend your current system, not replace it.

The side benefit from this perfect dove-tailing is that while Attribute Profiling is being implemented you can still provide your product recommendations through your standard means while there is insufficient Attribute Profile data. Once enough product and user Profile data has been established recommendations can then come from Attribute Profiling. In fact both systems can contribute at the same time to derive a product recommendation to a user. This symbiotic relationship can continue for as long as it is necessary to build the minimum data requirements. For example, the current system can be used where there is not enough data, such as a brand new user, and Attribute Profiling can take over once there are enough Attribute submissions to provide a more accurate recommendation service.

SUMMARY

In summary, the collaborative filtering industry has reached a plateau based around (what I refer to as) the genre-rating data model. This data model is comprised of organizing products around genre and allowing each user to apply a five-star rating to a particular product. While there are other data-models in the industry, such as Keywords and Tags, the genre-rating data model is the most successful and proliferated.

I liken the industry to a high-revving car that is in desperate need to shift gears. No matter how many tweaks you do to the engine you can only rev it so high! When all the cars are stuck in the same gear it only takes one to "shift" and blow away the competition.

Attribute Profiling is a shift in gears.

Attribute Profiling is a shift in gears.

As stated before, Attribute Profiling collects and organizes the necessary data to capture the "what" (what a product is like) and "why" (why a user likes/dislikes a product) while, simultaneously, providing a seamless way to compare and match the two together.

Focusing around adjectives, it captures a structured "fingerprint" for products and users' preferences that allow easy, apples-to-apples comparisons to identify accurate product recommendations.

Once implemented, the easy, intuitive and extended avenue in which a user can express their feelings will pull them to submit Attribute data for a given product. Also, as data increases, so too will the product recommendation accuracy, increasing the user experience. As the user experience increases via Attribute Profiling it will incite a stronger pull to submit more Attribute data (knowing that it is used to improve the recommendations).

This upward, self-feeding spiral will reach a critical point and take off very much like Google's search engine.

After all, a superior service is undoubtedly the most attractive beacon for any user.

Next steps

If you would like to learn more about Attribute Profiling please read the corollary white paper *Attribute Profiling: A complete walk-through*.

In addition to a more technical breakdown, it covers other features of Attribute Profiling such as:

- Sub-Profiles and how to leverage your existing data-model
- User Attribute Profiles that evolve as the user's preferences change
- Star-rated Attribute Profiles and the open possibilities