Chapter 4
The Long Tail in Recommender Systems

4.1 Introduction

The Long Tail is composed of a small number of popular items, the well-known *hits*, and the rest are located in the heavy tail, those not sell *that well*. The Long Tail offers the possibility to explore and discover—using automatic tools; such as recommenders or personalised filters—vast amounts of data. Until now, the world was ruled by the *Hit or Miss* categorisation, due in part to the shelf space limitation of the brick-and-mortar stores. A world where a music band could only succeed selling millions of albums, and touring worldwide.

Nowadays, we are moving towards the *Hit vs. Niche* paradigm, where there is a large enough availability of choice to satisfy even the most *Progressive–obscure–Spanish–metal* fan. The problem, though, is to filter and present the *right* artists to the user, according to her musical taste.

Chris Anderson introduces in his book, "The Long Tail" [1], a couple of important conditions to exploit the content available in niche markets. These are: (i) make everything available, and (ii) help me find it. It seems that the former condition is already fulfilled; the distribution and inventory costs are nearly negligible. Yet, to satisfy the latter we need recommender systems that exploit the *from hits to niches* paradigm. The main question, though, is whether current recommendation techniques are ready to assist us in this discovery task, providing recommendations of the *hidden gems* in the Long Tail.

In fact, recommenders that appropriately discount popularity may increase total sales, as well as potentially increase the margins by suggesting more novel, or less known, products [2]. Tucker et al. develop a theoretical model which shows how the existence of popular items can, in fact, benefit the perceived quality of niche products [3]. As these niche items are less likely to attract customers, the ones they attract perceive the products as higher quality than the mainstream ones. The authors' findings contribute to the understanding that popularity affects the long tail of e-Commerce. Even though web 2.0 tools based on the user's history of purchases promote the popular goods, their results suggest that mainstreamness benefits the

DOI 10.1007/978-3-642-13287-2_4,

perceived quality of niche products. Again, the big problem is to develop filters and tools that allow users to find and discover these niche products.

4.1.1 Pre- and post-filters

In the brick-and-mortar era, the market pre-filtered those products with lower probability of being bought by people. The main problem was the limited physical space to store the goods. Nowadays, with the unlimited shelf space, there is no need to pre-filter any product [1]. Instead, what users need are post-filters to make the products available and visible, and get personalised recommendations, according to their interests. Still, when publishers or producers pre-filter the content they also contribute to cultural production. E.g. many books or albums would be a lot worse without their editors and producers.

One should assume that there are some extremely poor quality products along the Long Tail. These products do not need to be removed by the gatekeepers anymore, but can remain in the Long Tail forever. The advisors are the ones in charge of not recommending low quality goods. In this sense, [4] proved that increasing the strength of social influence increased both inequality and unpredictability of success. As a consequence, popularity was only partly determined by quality. In fact, the quality of a work cannot be assessed in isolation, because our experience is so tied up with other people's experience of that work. Therefore, one can find items to match anyone's taste along the Long Tail. It is the job of the post-filters to ease the task of finding them.

4.2 The Music Long Tail

As already mentioned in Chap. 1, the "State of the Industry" report [5] presents some insights about the long tail in music consumption. For instance, 844 million digital tracks were sold in 2007, but only 1% of all digital tracks—the head part of the curve—accounted for 80% of all track sales. Also, 1,000 albums accounted for 50% of all album sales, and 450,344 of the 570,000 albums sold were purchased less than 100 times. Music consumption is biased towards a few mainstream artists. Ideally, by providing personalised filters and discovery tools to the listeners, music consumption would be diversified.

4.2.1 The Long Tail of Sales Versus the Long Tail of Plays

When computing a Long Tail distribution, one should define how to measure the popularity of the items. In the music domain, this can be achieved using the total number of sales or the total number of plays. On the one hand, the total number of sales denote the current trends in music consumption. On the other hand, the total number of playcounts tell us what people listen to, independently of the release year of the album (or song).

In terms of coverage, total playcounts is more useful, as it can represent a larger number of artists. An artist does not need to have an album released, but a *Myspace*-like page, which includes the playcounts for each song. Gathering information about the number of plays is easier than collecting the albums an artist has sold. Usually, the number of sales are shown in absolute values, aggregating all the information, and these numbers are used to compare the evolution of music consumption over the years. The total number of plays give us more accurate information, as it describes what people listen to. Thus, we will define the Long Tail in music using the total playcounts per artist.

As an example, Table 4.1 shows the overall most played artists at *last.fm* in July, 2007. These results come from more than 20 million registered users. Although the list of top-10 artists are biased towards this set of users, it still represents the listening habits of a large amount of people. In contrast, Table 4.2 shows the top-10

1.	The Beatles	(50,422,827)
2.	Radiohead	(40,762,895)
3.	System of a Down	(37,688,012)
4.	Red Hot Chili Peppers	(37,564,100)
5.	Muse	(30,548,064)
6.	Death Cab for Cutie	(29,335,085)
7.	Pink Floyd	(28,081,366)
8.	Coldplay	(27,120,352)
9.	Nine Inch Nails	(24,095,408)
10.	Blink 182	(23,330,402)

Table 4.1 Top-10 popular artists in *last.fm* according to the total number of plays (last column). Data gathered during July, 2007.

artists in 2006 based on total digital track sales (last column) according to Nielsen Soundscan 2006 report [6]. The second column (values in parenthesis) shows the corresponding *last.fm* artist rank. There is not a clear correlation between the two lists, and only one artist (*Red Hot Chili Peppers*) appears in both top-10 lists.

Furthermore, Table 4.3 shows the top-10 selling artists in 2006 based on total album sales (last column), again according to the Nielsen 2006 report. In this case, classic artists such as *Johnny Cash* (top-2) or *The Beatles* (top-5) appear. This reflects the type of users that still buy CDs. Regarding *Carrie Underwood*, she is an American country pop music singer who became famous after winning the fourth season of *American Idol* (2005). *Carrie Underwood* album, released in late 2005, became the fastest selling debut Country album. *Keith Urban*, *Tim McGraw* and *Rascal Flatts* are American country/pop songwriters with a leading male singer. In all these cases, they are not so popular in the *last.fm* community.

All in all, only *The Beatles* (in Table 4.3), and *Red Hot Chili Peppers* (in Table 4.2) appear in the top-10 *last.fm* chart (see Table 4.1). It is worth noting that in 2006 *The Beatles* music collection was not (legally) available for purchase in digital

1.	(912)	Rascal Flatts	(3,792,277)
2.	(175)	Nickelback	(3,715,579)
3.	(205)	Fray	(3,625,140)
4.	(154)	All-American Rejects	(3,362,528)
5.	(119)	Justin Timberlake	(3,290,523)
6.	(742)	Pussycat Dolls	(3,277,709)
7.	(4)	Red Hot Chili Peppers	(3,254,306)
8.	(92)	Nelly Furtado	(3,052,457)
9.	(69)	Eminem	(2,950,113)
10.	(681)	Sean Paul	(2,764,505)

Table 4.2 Top-10 artists in 2006 based on total digital track sales (last column) according to Nielsen report. The second column (values in parenthesis) shows the corresponding *last.fm* artist rank.

1.	(912)	Rascal Flatts	(4,970,640)
2.	(70)	Johnny Cash	(4,826,320)
3.	(175)	Nickelback	(3,160,025)
4.	(1514)	Carrie Underwood	(3,016,123)
5.	(1)	The Beatles	(2,812,720)
6.	(1568)	Tim McGraw	(2,657,675)
7.	(2390)	Andrea Bocelli	(2,524,681)
8.	(1575)	Mary J. Blige	(2,485,897)
9.	(1606)	Keith Urban	(2,442,577)
10.	(119)	Justin Timberlake	(2,437,763)

Table 4.3 Top-10 selling artists in 2006 (based on total album sales, last column) according to Nielsen report. The second column (values in parenthesis) shows the corresponding *last.fm* artist rank.

form. On the other hand, *last.fm* listening habits denote what people listen to, and that does not necessarily correlate with the best sellers. For instance, classic bands such as *Pink Floyd*, *Led Zeppelin* (at top-15), *Tool* (top-16) or *Nirvana* (top-18) did not release any new album during 2006, but still they are in the top-20 (at mid-2007). From this informal analysis we conclude that popularity is a nebulous concept that can be viewed in different ways.

From now on, we characterise music popularity using the total playcounts of an artist, keeping in mind that the data is not correlated with the actual number of sales, and also that the data will be biased towards the subset of users that are taken into account (in our case, the entire *last.fm* community).

4.2.2 Collecting Playcounts for the Music Long Tail

In the music field, total artist playcounts allow us to determine artist popularity. There are at least two different ways to collect artists' plays from the web. The first one is using *last.fm* data, and the second one is using the data from *Myspace*. In principle, one should expect a clear correlation among both datasets. That is, if an artist has a lot of plays in one system then the same should happen in the other one.

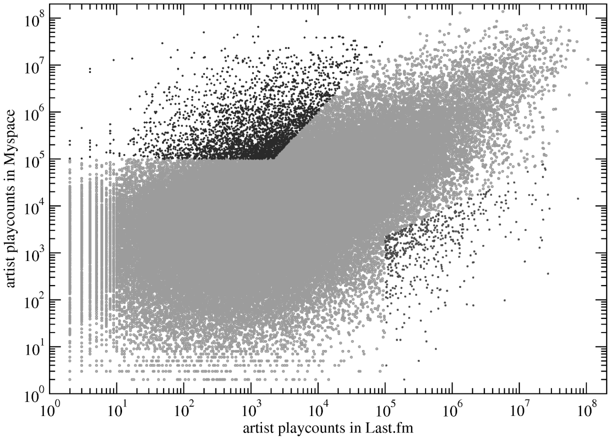

Fig. 4.1 Correlation between *last.fm* and *Myspace* artist playcounts. Data gathered during January, 2008.

However, each system measures different listening habits. On the one hand, *last.fm* monitors what users listen to in virtually any device, whereas *Myspace* only tracks the number of times a song has been played in their embedded Flash player. On the other hand, *Myspace* data can track the number of plays for those artists that have not released any album, but a list of songs (or demos) that are available on the *Myspace* artist profile. In this case, it is very unlikely to gather this data from *last.fm* because the only available source to listen to the songs is via *Myspace* (specially if the artist forbids users to download the songs from *Myspace*). For example, the artist *Thomas Aussenac* has (on October 21st, 2008) 12,486 plays in *Myspace*[1] but only 63 in *last.fm*.[2] Therefore, sometimes (e.g. head and mid artists) both systems can provide similar listening habits results, whilst in other cases they track and measure different trends. Some plausible reasons about these differences could be due to the demographics and locale of both users and artists in the two systems.

Figure 4.1 depicts the total playcounts for an artist in *last.fm* versus the total playcounts in *Myspace* (data gathered during January, 2008). That is, given the playcounts of an artist in *last.fm*, it plots its total plays in *Myspace*. We remark two interesting areas; upper left and bottom right. These areas are the ones with those artists whose playcounts are clearly uncorrelated between the two datasets. For instance, the upper left area shows the artists that have lots of plays in *Myspace*, but just a few in *last.fm*. The formula used to select the artists in this area is (it is analogous for the *last.fm* versus *Myspace*—in the bottom right area):

$$Plays_{Myspace} > 10^5 \wedge \frac{log(Plays_{Myspace})}{log(Plays_{Last.fm})} \geq 1.5 \tag{4.1}$$

[1] `http://www.myspace.com/thomasaussenac`

[2] `http://www.last.fm/music/thomas+aussenac`

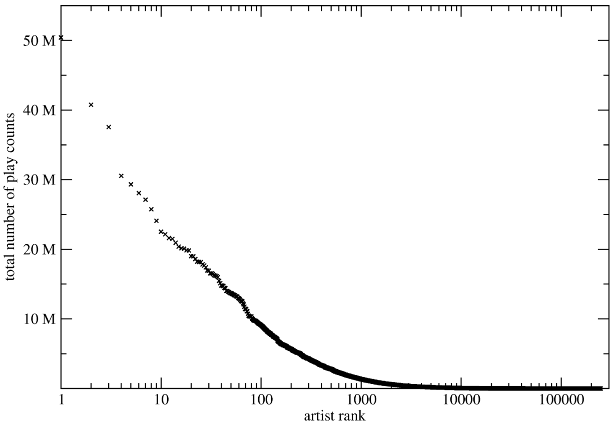

Fig. 4.2 The music Long Tail effect. A log-linear plot depicting the total number of plays per artist. Data gathered during July, 2007, for a list of 260,525 artists.

That is, artists that have more than 100,000 plays in *Myspace*, but much less in *last.fm*. In this case, we could consider that some of these artists are well-known in the *Myspace* area, having lots of fans that support them, but the artist still has no effect outside *Myspace*. Maybe this type of artists can reach a broader popularity after releasing an album. For instance, *Michael Imhof*,[3] a German *house* and *r&b* artist, has more than 200,000 playcounts in *Myspace*, but only 2 in *last.fm*. A more extreme example is *Curtis Young*[4] (aka *Hood Surgeon*), the son of legendary hip-hop producer *Dr. Dre*, who has 13,814,586 plays in *Myspace* but less than 20,000 in *last.fm*. It is worth mentioning that there are some services[5] that allow a *Myspace* artist to automatically increase their total playcounts, without the need for real users.

All in all, there are different ways of measuring an artist's popularity, and might even exist different *domains* of popularity; what is popular in one domain can be unknown in another. As previously stated, popularity is a nebulous concept that can be viewed in different ways.

4.2.3 An Example

Figure 4.2 depicts the Long Tail popularity, using total playcounts, for 260,525 music artists. The horizontal axis contains the list of artists ranked by its total playcounts. For example *The Beatles*, at position 1, has more than 50 million playcounts.

This data was gathered from *last.fm* during July, 2007. *Last.fm* provides plugins for almost any desktop music player (as well as *iPhones* and other mobile

---

[3] `http://www.myspace.com/michaelimhof`

[4] `http://www.myspace.com/curtisyoungofficial`

[5] Such as `http://www.somanymp3s.com/`

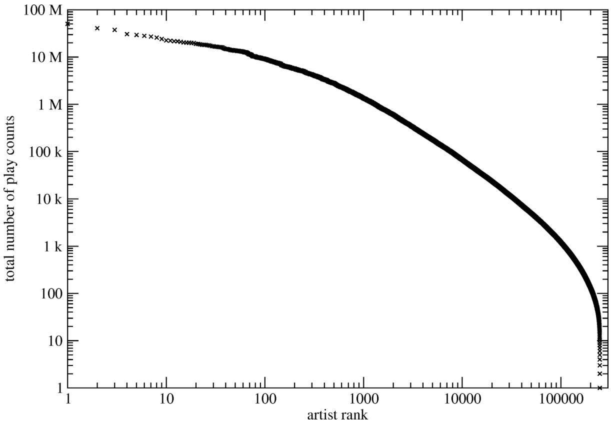

Fig. 4.3 The music Long Tail effect. Same plot as Fig. 4.2 here in log–log scale. The best fit is a log-normal distribution, with a mean of log $\mu = 6.8$, and standard deviation of log, $\sigma = 2.18$. The fast drop in the tail is in part due to misspelled artists (e.g. incorrect metadata in the ID3 tags).

devices) to track users' listening behaviour. It also provides a Flash player embedded in their website, and a client for PC, Mac and Linux that can create personalised audio streams. Figure 4.2 corroborates the music consumption reports by Nielsen Soundscan [5]; a few artists concentrate most of the total plays, whilst many musicians hold the rest. Figure 4.3 presents the same data as Fig. 4.2, in log–log scale. The best fit for the curve is a log-normal distribution, with parameters mean of log $\mu = 6.8$, and standard deviation of log $\sigma = 2.18$ (more information about fitting a curve with a distribution model is presented in Sec. 4.3.2). It is worth noting that the fast drop in the tail is in part due to misspelled artists (e.g. incorrect metadata in the ID3 tags).

4.3 Definitions

The Long Tail of a catalog is measured using the frequency distribution (e.g. purchases, downloads, etc.), ranked by item popularity. We present now two definitions for the Long Tail. The first one is an informal, intuitive one. The second one is a quantitative definition that uses a formal model to characterise the shape of the curve, and a method to fit the data to some well-known distributions (e.g. power-law, power-law with exponential decay, log-normal, etc.).

4.3.1 Qualitative, Informal Definition

According to Chris Anderson [1], the Long Tail is divided in two separate parts: the head and the tail. The head part contains the items one can find in the *old* brick-and-mortar markets. The tail of the curve is characterised by the remainder of the existing products. This includes the items that are available in on-line markets. Chris Anderson's definition, based on the economics of the markets, is:

> *The Long Tail is about the economics of abundance; what happens when the bottlenecks that stand between supply and demand in our culture start to disappear and everything becomes available to everyone.*

The definition emphasises the existence of two distinguished markets; the familiar one (the *Head*), and the long ignored but emerging since the explosion of the web (the *Tail*), consisting of small niche markets.

Another definition is the one by Jason Foster:

> *The Long Tail is the realization that the sum of many small markets is worth as much, if not more, than a few large markets.*[6]

Both definitions are based on markets and economics, and do not propose any computational model to compute and characterise any tail curve, nor fit the data to any existing distribution. Indeed, [1] does not define how to split the head and the tail parts, that are the two key elements in both definitions.

4.3.1.1 Physical Apples Versus Online Oranges

Since *The Long Tail* book became a top-seller, there is a lot of criticism against Anderson's theory. The most common criticism is the lack of scientific backup when comparing different data sources. That is, when comparing the online world to the physical world, Anderson simplifies too much. For instance, he considers only one brick-and-mortar store (e.g. *Walmart*), and compares their music catalog with the one found in the *Rhapsody* online store. However, in the real world there are much more music stores than *Walmart*. Indeed, there are specialised music stores that carry out ten times the volume of *Walmart*'s music catalog. Sadly enough, these ones are completely ignored in Anderson's studies [7].

In addition, there is no clear evidence that online stores can monetise the Long Tail. According to Elberse et al. there is no evidence of a shift in online markets towards promoting the tail [8]. The tail is long, but extremely flat. In their results, hit-driven economies are found in both physical and online markets. Furthermore, in an older study [9], Elberse found that the long tail of movies, those that sell only a few copies every week nearly doubled during their study period. However, the number of non-selling titles rose four times, thus increasing the size of the tail. Regarding the head of the curve; a few mainstream movies still accounted for most of the sales.

---

[6] From `http://longtail.typepad.com/the_long_tail/2005/01/definitions_fin.html`

Another drawback of the theory is the creation of online oligarchies. “Make everything available” is commonly achieved by *One-Big-Virtual-Tent* rather than *Many-Small-Tents*.[7] That is to say, there is only one *Amazon* that provides most of the content.

Last but not least, Anderson’s theory states that the Long Tail follows a power-law distribution. That is a straight line in a log–log plot. However, only plotting a curve in a log–log scale is not enough to verify that the curve follows a power-law. It can better fit to other distributions, such as log-normal or a power-law with an exponential decay of the tail. We need, then, a model that allows us to quantitative define the shape of the Long Tail curve, without the need of linking it with niche markets, economics, and profitable (or not) e-Commerce websites.

4.3.2 Quantitative, Formal Definition

The Long Tail model, $F(x)$, simulates any heavy-tailed distribution [10]. It models the cumulative distribution of the Long Tail data. $F(x)$ represents the share (%) of total volume covered by objects up to rank x:

$$F(x) = \frac{\beta}{(\frac{N_{50}}{x})^{\alpha} + 1} \tag{4.2}$$

where α is the factor that defines the S-shape of the function, β is the total volume share (and also describes the amount of latent demand), and N_{50}, the median, is the number of objects that cover half of the total volume, that is $F(N_{50}) = 50$.

Once the Long Tail is modelled using $F(x)$, we can divide the curve in three parts: head, mid, and the tail. The boundary between the head and the mid part of the curve is defined by:

$$X_{head \to mid} = N_{50}^{2/3} \tag{4.3}$$

Likewise, the boundary between the mid part and the tail is:

$$X_{mid \to tail} = N_{50}^{4/3} \simeq X_{head \to mid}^{2} \tag{4.4}$$

Figure 4.4 depicts the cumulative distribution of the Long Tail of the 260,525 music artists presented in Fig. 4.2. Interestingly enough, the top-737 artists, 0.28% of all the artists, account for 50% of the total playcounts, $F(737) = 50 (N_{50} = 737)$, and only the top-30 artists hold around 10% of the plays. Another measure is the *Gini coefficient*. This coefficient measures the inequality of a given distribution, and it determines the degree of imbalance [11]. In our Long Tail example, 14% of the artists hold 86% of total playcounts, yielding a Gini coefficient of 0.72. This value describes a skewed distribution, higher than the classic 80/20 Pareto rule, with a

[7] See Tom Slee critical reader’s companion to “The Long Tail” book at `http://whimsley.typepad.com/whimsley/2007/03/the_long_tail_l.html`

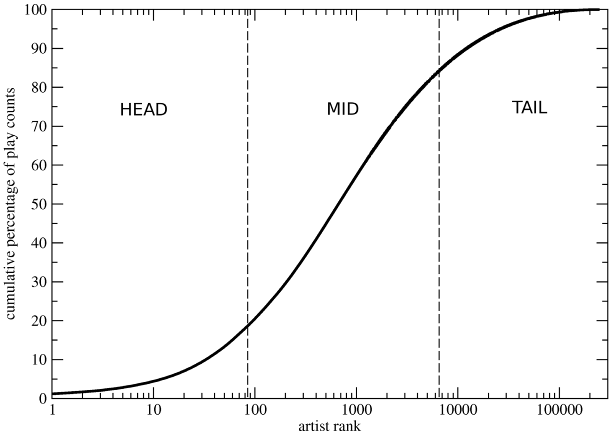

Fig. 4.4 Example of the Long Tail model. It shows the cumulative percentage of playcounts of the 260,525 music artists from Fig. 4.2. Only top-737 artists, 0.28% of all the artists, accumulates the 50% of total playcounts (N_{50}). Also, the curve is divided in three parts: head, mid and tail ($X_{head \to mid} = 82$, and $X_{mid \to tail} = 6,655$), so each artist is located in one section of the curve.

value of 0.6. Figure 4.4 also shows the three different sections of the Long Tail. The head of the curve, $X_{head \to mid}$ consists of only 82 artists, whilst the mid part has 6,573 ($X_{mid \to tail} = 6,655$). The rest of the artists are located in the tail.

4.3.2.1 Fitting a Heavy-Tailed Distribution Using $F(x)$

To use the $F(x)$ function we need to fit the curve with an estimation of α, β and N_{50} parameters. We do a non-linear regression, using Gauss–Newton method for non-linear least squares, to fit the observations of the cumulative distribution to $F(x)$.[8] Figure 4.5 shows an example of the fitted distribution using the $F(x)$ model. The data is the one from artist popularity in *last.fm* (Fig. 4.4).

4.3.3 Qualitative Versus Quantitative Definition

On the one hand, the qualitative definition by Anderson emphasises the economics of the markets, and the shift from physical to virtual, online, goods. On the other hand, the quantitative definition is based on a computational model that allows us to fit a set of observations (of the cumulative distribution) to a given function, $F(x)$.

The main difference between the two definitions (qualitative and quantitative) is the way each method split the curve into different sections (e.g. the head and

[8] To solve the non-linear least squares we use the *R* statistical package. The code is available at `http://mtg.upf.edu/~ocelma/PhD`

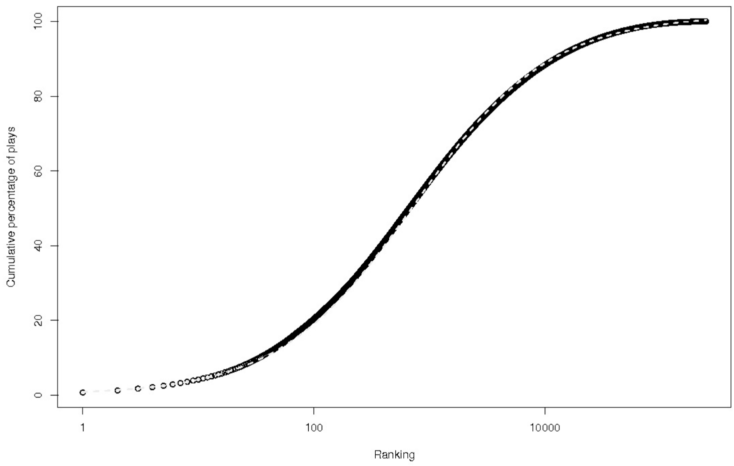

Fig. 4.5 Example of fitting a heavy-tailed distribution (the one in Fig. 4.4) with $F(x)$. The *black dots* represent the observations while the *white dotted curve* represents the fitted model, with parameters $\alpha = 0.73$, and $\beta = 1.02$.

the tail). The qualitative approach is based on the % covered by x (e.g. "20% of the products represent 80% of sales") whereas the quantitative definition splits the x (log) axis equally in three (head, mid, and tail) parts. The main problem is that when adding many more products (say 10,000) in the curve, the changes in the *head* and *tail* boundaries are very radical in the qualitative definition. The quantitative approach does not suffer from this problem. The changes in the section boundaries are not so extreme.

4.4 Characterising a Long Tail Distribution

An early mention of the "long tail", in the context of the Internet, was Clay Shirky's essay in February, 2003.[9] After that, [1] converted the term to a proper noun, and defined a new trend in economics. Since then, the spotlight on the "Long Tail" noun has created many different opinions about it.

In our context, we use a "Long Tail" curve to describe the popularity phenomenon in any recommender system, to show how popularity can affect the recommendations. So, given a long tail distribution of the items' popularity, an important step is to characterise the shape of the curve to understand the amount of skewness.

[9] See `http://shirky.com/writings/powerlaw_weblog.html`

We characterise a Long Tail distribution using Kilkki's $F(x)$ function. Its parameters α, β, and N_{50} defines the shape of the curve.

Yet, it is also important to determine the shape of the curve according to well-known probability density distributions. There are different probability density distribution functions that can fit a heavy-tailed curve. We present some of them here: power-law, power-law with exponential decay, and log-normal distribution.

4.4.1 Not All Long Tails Are Power-Law

A *power-law distribution* is described using the probability density distribution (*pdf*), $f(x)$:

$$f(x) = ax^{-\gamma} \tag{4.5}$$

Power-law distribution has the property of (asymptotic) scale invariance. This type of distribution cannot be entirely characterised by its mean and variance. Also, if the γ power-law exponent has a value close to 1, $\gamma \simeq 1$, then this means that the long tail is fat.[10] In other words, a power-law with $\gamma \gg 1$ consists of a thin tail (with values close to 0), and a short head with a high probability value.

Power-law with an exponential decay distribution differs from a power-law by the shape of the tail. Its *pdf* is defined by:

$$f(x) = x^{-\gamma}e^{-\lambda x}, \tag{4.6}$$

There exists an N that denotes the threshold between the power-law distribution ($x^{-\gamma}, x \leq N$), and the exponential decay ($e^{-\lambda x}, x > N$). This means that the tail of the curve is better represented with an exponential cut-off.

In a *log-normal* distribution the logarithm of the variable is normally distributed. That is to say, if a variable X is normally distributed, then $Y = e^X$ has a log-normal distribution. Log-normal distribution promotes the head of the curve. It is a distribution skewed to the right, where the popular items have a strong effect, whilst the tail has a very small contribution in the *pdf*:

$$f(x) = \frac{1}{x} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \tag{4.7}$$

Thus, the main problem is, given a curve in a log–log scale representation, to decide which is the best model that explains the curve. It is worth noting that, according to Anderson's theory (i.e. the Long Tail is profitable), the curve should be modelled as a power-law, with $\gamma \simeq 1$, meaning that the tail is fat. However, if the best fit is using another distribution, such as a log-normal—which is very common— then Anderson's theory cannot be strictly applied in that particular domain, and context.

[10] This is the only case where Anderson's Long Tail theory can be applied.

4.4.2 A Model Selection: Power-Law or Not Power-Law?

To characterise a heavy-tailed distribution, we follow the steps described in Clauset et al. [12]. As previously mentioned, the main drawbacks when fitting a Long Tail distribution are: (i) to plot the distribution on a log–log plot, and see whether it follows a straight line or not, and (ii) use linear regression by least squares to fit a line in the log–log plot, and then use R^2 to measure the fraction of variance accounted for the curve. This approach gives a poor estimate of the model parameters, as it is meant to be applied to regression curves, not to compare distributions. Instead, to decide whether a heavy-tailed curve follows a power-law distribution, [12] propose the following steps:

1. *Estimate* γ. Use the maximum likelihood estimator (MLE) for the γ scaling exponent. MLE always converge to the correct value of the scaling exponent.
2. *Detect* x_{min}. Use the goodness of fit value to estimate where the scaling region begins (x_{min}). The curve can follow a power-law on the right or upper tail, so above a given threshold x_{min}. The authors propose a method that can empirically find the best scaling region, based on the Kolmogorov–Smirnov D statistic.
3. *Goodness of the model*. Use, again, the Kolmogorov–Smirnov D statistic to compute the discrepancy between the empirical distribution and the theoretical one. The Kolmogorov–Smirnov (K–S) D statistic will converge to zero, if the empirical distribution follows the theoretical one (e.g. power-law). The K–S D statistic for a given cumulative distribution function $F(x)$, and its empirical distribution function $F_n(x)$ is:

$$D_n = \sup_x |F_n(x) - F(x)|, \tag{4.8}$$

   where $\sup |S|$ is the supremum of a set S. That is the lowest element of S that is greater than or equal to each element of S. The supremum is also referred to as the *least upper bound*.
4. *Model selection*. Once the data is fitted to a power-law distribution, the only remaining task is to check among the different alternatives. That is, to detect whether other non power-law distributions could have produced the data. This is done using pairwise comparison (e.g. power-law versus power-law with exponential decay, power-law versus a log-normal, etc.), and [12] use the Vuong's test [13]. Vuong's test uses the log-likelihood ratio and the Kullback–Leibler information criterion to make probabilistic statements about the two models. Vuong's statistical test is used for the model selection problem, where one can determine which distribution is closer to the real data. A large, positive Vuong's test statistic provides evidence of the best fitting using a power-law distribution over the other distribution, while a large, negative test statistic is an evidence of the contrary.

4.5 The Dynamics of the Long Tail

Another important aspect of any Long Tail is its dynamics. E.g., does an artist stay in the head region forever? Or the other way around; will niche artists always remain in the long tail? Figure 4.6 depicts the increase of the Long Tail popularity after 6 months, using 50,000 out of the 260,525 *last.fm* artists (see Fig. 4.2). Figure 4.6 shows the dynamics of the curve comparing two snapshots; one from July 2007, and the other from January 2008. The most important aspect is the increase of total playcounts in each area of the curve.

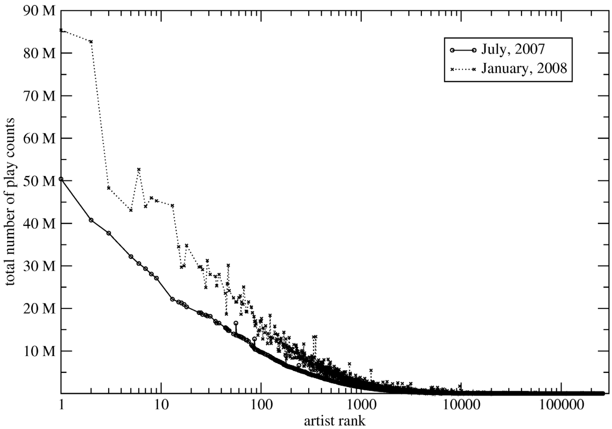

Fig. 4.6 The dynamics of the Long Tail after 6 months (between July, 2007 and January, 2008). *Radiohead*, at top-2, is now closer to *The Beatles* (top-1), due to the release of their *In Rainbows* album.

4.5.1 Strike a Chord?

Table 4.4 shows the playcount increment, in %. In all the three regions—head, mid, and tail—the percentage increment of plays is almost the same (around 62%), meaning that not many artists move between the regions. For instance, in the head area, *Radiohead* at top-2 is much closer to top-1, *The Beatles*, due to the release of the *In Rainbows* album.[11] Still, the band remains at top-2. An interesting example in the tail area is the *Nulla Costa* band. This band was at rank 259,962 in July, 2007. After 6 months they increase from 3 *last.fm* playcounts to 4,834, positioning at rank 55,000. Yet, the band is still in the tail region. We could not detect any single artist

[11] *In Rainbows* album was released on October 10th, 2007

that clearly moved from the tail to the mid region.[12] There exist niche artists, and the main problem is to find them. The only way to leverage the long tail is by providing recommendations that promote unknown artists.

Once the Long Tail is formally described, the next step is to use this knowledge when providing recommendations. The following section presents how one can exploit the Long Tail to provide novel or familiar recommendations, taking into account the user profile.

Long Tail region	Increase (%)
Head	61.20
Mid	62.29
Tail	62.32

Table 4.4 Increase of the Long Tail regions (in %) after 6 months (comparing two snapshots in July, 2007 and January, 2008).

4.6 Novelty, Familiarity and Relevance

If you like The Beatles you might like...X. Now, ask several different people and you will get lots of different $X's$. Each person, according to her ties with the band's music, would be able to propose interesting, surprising or expected $X's$. Nonetheless, asking the same question to different recommender systems we are likely to get similar results. Indeed, two out of five tested music recommenders contain John Lennon, Paul McCartney and George Harrison in their top-10 (*last.fm* and *the.echotron.com* by *The Echo Nest* company). *Yahoo! Music* recommends John Lennon and Paul McCartney (1st and 4th position), whereas *Mystrands.com* only contains John Lennon (at top-10). Neither *ilike* nor *Allmusic.com* contain any of these musicians in their list of *Beatles*' similar artists. Furthermore, Amazon's top-30 recommendations for the *Beatles*' *White Album* is strictly made of other *Beatles*' albums (all of a sudden, at the fourth page of the navigation there is the first non-*Beatles* album; *Exile on Main St.* by The Rolling Stones). Finally, creating a playlist from *OneLlama.com*—starting with a *Beatles* seed song—one gets four out of ten songs from the *Beatles*, plus one song from John Lennon, so it makes half of the playlist. It is worth mentioning that these recommenders use different approaches, such as: collaborative filtering, social tagging, web mining and co-occurrence analysis of playlists. To conclude this informal analysis, the most noticeable fact is that only *last.fm* remembers Ringo Starr![13]

[12] *Last.fm* has the "hype artist" weekly chart, `http://www.last.fm/charts/hypeartist`, a good source to track the movements in the Long Tail curve.

[13] This informal analysis was done in July, 2007.

One can agree or disagree with all these *Beatles*' similar artist lists. However, there are a very few, if none at all, serendipitous recommendations (the rest of the similar artists were, in no particular order: *The Who*, *The Rolling Stones*, *The Beach Boys*, *The Animals*, and so on). Indeed, some of the before mentioned systems provide filters, such as: "surprise me!" or the "popularity slider", to dive into the Long Tail of the catalog. Thus, novel recommendations are sometimes necessary to improve the user's experience and discovery in the recommendation workflow.

It is not our goal to decide whether one can monetise the Long Tail or to exploit the niche markets, but to help people discover those items that are lost in the tail. Hits exist and they always will. Our goal is to motivate and guide the discovery process, presenting to users rare, non-hit, items they could find interesting.

4.6.1 Recommending the Unknown

It has been largely acknowledged that item popularity can decrease user satisfaction by providing obvious recommendations [14, 15]. Yet, there is no clear recipe for providing *good* and *useful* recommendations to users. We can foresee at least three key aspects that should be taken into account. These are: novelty and serendipity, familiarity, and relevance [16]. According to Wordnet dictionary,[14] **novel** (*adj.*) has two senses: "new—original and of a kind not seen before"; and "refreshing—pleasantly new or different". Serendipity (*noun*) is defined as "good luck in making unexpected and fortunate discoveries". Familiar (*adj.*) is defined as "well known or easily recognised". In our context, we measure the novelty for a given user u as the ratio of unknown items in the list of top-N recommended items, $\mathcal{L}_N$:

$$Novelty(u) = \frac{\sum_{i \in \mathcal{L}_N} (1 - Knows(u,i))}{N}, \tag{4.9}$$

being $Knows(u,i)$ a binary function that returns 1 if user u already knows item i, and 0 otherwise. Likewise, user's familiarity with the list of recommended items can be defined as $Familiar(u) = 1 - Novelty(u)$.

Nonetheless, a user should be familiar with some of the recommended items, to improve confidence and trust in the system. Also, some items should be unknown to the user (discovering *hidden* items in the catalog). A system should also give an explanation of why those—unknown—items were recommended, providing a higher confidence and transparency on these recommendations. The difficult job for a recommender is, then, to find the proper level of familiarity, novelty and relevance for *each* user.

Figure 4.7 shows the long tail of item popularity, and it includes a user profile. The profile is exhibited as the number of times the user has interacted with that item. Taking into account item popularity plus the user profile information, a

[14] `http://wordnet.princeton.edu`

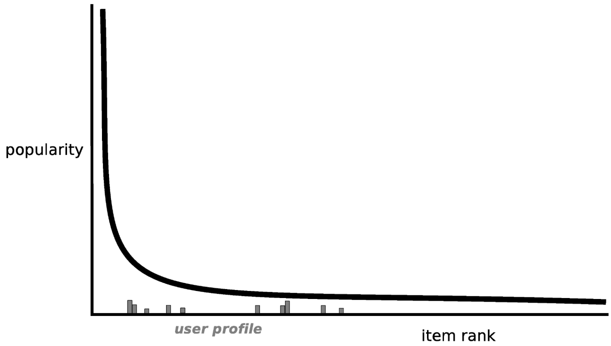

Fig. 4.7 A user profile represented in the Long Tail. The profile is exhibited as the number of times the user has interacted with that item.

recommender can provide personalised and relevant recommendations that are also novel to the user.

4.6.1.1 Trade-Off Between Novelty and Relevance

However, there is a trade-off between novelty and user's relevance. The more novel, unknown items a recommender presents to a user, the less relevant they can be perceived by her.

Figure 4.8 presents the trade-off between novelty and relevance. It shows the different recommendation states for a given a user *u*, given a large collection of items (say, not only the user's personal collection). The gray triangle represents the area where a recommender should focus on to provide relevant items to *u*. On the one hand, laid-back recommendations (bottom-right) appear when the system recommends familiar and relevant items to *u*. On the other hand, the discovery process (top-right) starts when the system provides to the user (potentially) unknown items that could fit in her profile. The provided recommendations should conform to the user's intentions; sometimes a user is expecting familiar recommendations (laid-back state), while in other cases she is seeking to actively discovery new items.

There are two more cases, that is when the recommender provides popular items, and when it provides random ones. This can happen when there is not enough information about the user (e.g. the user cold-start problem). In this case, the system can recommend popular items (bottom-left). Popular items are expected to be somehow familiar to the user, but not necessarily relevant to her. The other situation is when the system provides random recommendations to *u* (top-left). This case is similar to a shuffle playlist generator, with the difference that in our case the items' catalog is much bigger than the personal music collection of *u*. Thus, there is less chances that user *u* might like any of the random recommendations, as they are not personalised at all.

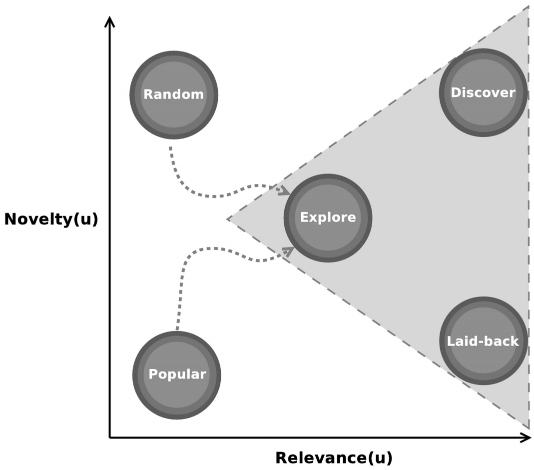

Fig. 4.8 Trade-off between novelty and relevance for a user *u*.

4.6.2 Related Work

Serendipity and novelty are relevant aspects in the recommendation workflow [15]. Indeed, there is some related work that explicitly addresses these aspects. For instance, five measures to capture redundancy are presented in [17]. These measures allow one to infer whether an item—that is considered relevant—contains any novel information to the user. Yang and Li [18] defines novelty in terms of the user knowledge and her degree of interest in a given item. In [19], Weng et al. propose a way to improve the quality and novelty of the recommendations by means of a topic taxonomy-based recommender, and hot topic detection using association rules. Other proposals include disregarding items if they are too similar to other items that the user has already seen [20], or simple metrics to measure novelty and serendipity based on the average popularity of the recommended items [21].

Even though all these approaches focus on providing novel and serendipitous recommendations, there is no framework that consistently evaluates the provided recommendations. Thus, there is a need to design evaluation metrics to deal with the effectiveness of novel recommendations, not only measuring prediction accuracy, but taking into account other aspects such as usefulness and quality [14, 22]. Novelty metrics should look at how well a recommender system made a user aware of previously unknown items, as well as to what extent users accept the new recommendations [14].

Generally speaking, the most popular items in the collection are the ones with higher probability that a given user will recognise, or be broadly familiar with. Likewise, one can assume that items with less interaction—rating, purchasing, previewing—within the community of users are more likely to be unknown [21]. In this sense, the Long Tail of the items' catalog assists us in deciding how novel or familiar an item could be. Yet, a recommender system must predict whether an item could be relevant, and then be recommended, to a user.

4.7 Summary

Effective recommendation systems should promote novel and relevant material (non-obvious recommendations), taken primarily from the tail of a popularity distribution. In this sense, the Long Tail can be described in terms of niche markets' economics, but also by describing the item popularity curve. We use the latter definition—the Long Tail model, $F(x)$—to describe the cumulative distribution of the curve. In the music field, the $F(x)$ model allows us to define artist popularity, and her location in the curve (head, mid or tail region). Hence, $F(x)$ denotes the shared knowledge about an artist, by a community of listeners. From this common knowledge, we can derive whether an artist can be novel and relevant to a given user profile.

Our results show that music listening habits follow the hit-driven (or mainstream) paradigm; 0.28% (737 out of 260,525) of the artists account for the 50% of total playcounts. The best fit (in the log–log plot) for the music Long Tail is a log-normal distribution. A log-normal distribution concentrates most of the information in the the head region. Even though we use playcounts and not total sales to populate the curve, this finding unveils Anderson's theory about the economics and monetisation in the Long Tail. Despite Anderson's failure or success theory, his core idea still is an interesting way to explain the changes the web has provoked, in terms of the availability of all kind of products—from hits to niches.

One of the goals of a recommender should be to promote the tail of the curve by providing relevant, personalised novel recommendations to its users. That is, to smoothly interconnect the head and mid regions with the tail, so the recommendations can drive interest from one to the other. Figure 4.9 presents this idea. It depicts a 3D representation of the Long Tail; showing the item popularity curve, the similarities among the items, and a user profile denoted by her preferred items (in dark gray colour). The set of candidate items (dotted lines) to be recommended to the user are shown also. Items' height denotes the relevance for that user. Candidate items located in the tail part are considered more novel—and, potentially relevant—than the ones in the head region.

4.7.1 Links with the Following Chapters

In this chapter we have presented the basics for novelty detection in a recommender system, using the popularity information and its Long Tail shape. The next step is to evaluate these types of recommendations. We can foresee two different ways to evaluate novel recommendations, and these are related with (i) exploring the available (and usually, very large) item catalog, and (ii) filtering new incoming items.

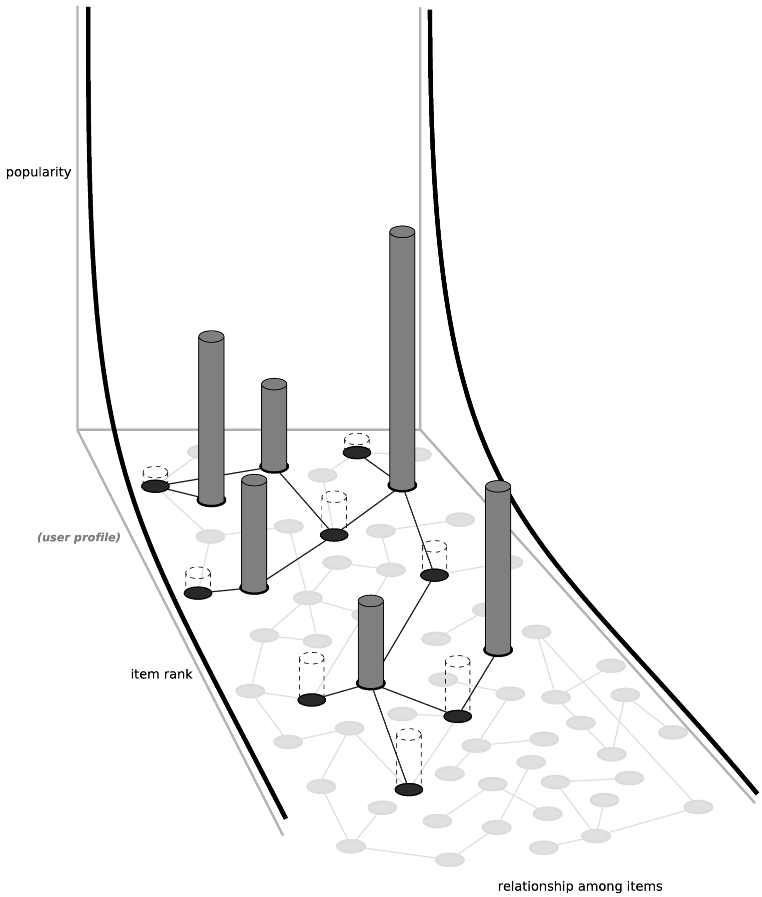

Fig. 4.9 A 3D representation of the Long Tail. It adds another dimension; the similarities among the items, including the representation of a user profile (*in gray*). The set of candidate items to be recommended to the user are shown (*in dotted lines*) and its height denotes the relevance for the user.

We mainly focus on the former case, and we present two complementary evaluation methods. On the one hand, *network-centric* evaluation method (presented in Chap. 6) focuses on analysing the items' similarity graph, created using any item-based recommendation algorithm. The aim is to detect whether the intrinsic topology of the items' network has any pathology that hinders novel recommendations, promoting the most popular items. On the other hand, a *user-centric* evaluation aims at measuring the perceived quality of novel recommendations. This user evaluation is presented in Chap. 7. Yet, before presenting the evaluation results we introduce, in Chap. 5, the metrics that we use.

References

1. C. Anderson, *The Long Tail. Why the Future of Business Is Selling Less of More*. New York, NY: Hyperion, 2006.
2. D. M. Fleder and K. Hosanagar, "Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity," *SSRN eLibrary*, 2007.
3. C. Tucker and J. Zhang, "How does popularity information affect choices? theory and a field experiment," *SSRN eLibrary*, 2008.
4. M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," *Science*, vol. 311, pp. 854–856, February 2006.
5. N. Soundscan, "State of the industry," *Nielsen Soundscan Report. National Association of Recording Merchandisers*, 2007.
6. N. Soundscan, "Nielsen soundscan report. Year–end music industry report," White Plains, NY, 2006.
7. T. Slee, "A critical reader's companion to the long tail," 2006.
8. A. Elberse, "Should you invest in the long tail?" *Harvard Business Review*, vol. 86, no. 7/8, pp. 88–96, 2008.
9. A. Elberse and F. Oberholzer-Gee, "Superstars and underdogs: An examination of the long tail phenomenon in video sales," *Harvard Business School Working Paper*, May 2006.
10. K. Kilkki, "A practical model for analyzing long tails," *First Monday*, vol. 12, May 2007.
11. C. Gini, "Measurement of inequality and incomes," *The Economic Journal*, vol. 31, pp. 124–126, 1921.
12. A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Reviews*, June 2007.
13. Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, vol. 57, pp. 307–333, March 1989.
14. J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transaction on Information System*, vol. 22, no. 1, pp. 5–53, 2004.
15. S. M. McNee, J. Riedl, and J. A. Konstan, "Being accurate is not enough: How accuracy metrics have hurt recommender systems," in *Computer Human Interaction. Human factors in computing systems*, (New York, NY), pp. 1097–1101, ACM, 2006.
16. O. Celma and P. Lamere, "Music recommendation tutorial," in *Proceedings of 8th International Conference on Music Information Retrieval*, (Vienna, Austria), 2007.
17. Y. Zhang, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," in *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (New York, NY), pp. 81–88, ACM, 2002.
18. Y. Yang and J. Z. Li, "Interest-based recommendation in digital library," *Journal of Computer Science*, vol. 1, no. 1, pp. 40–46, 2005.
19. L.-T. Weng, Y. Xu, Y. Li, and R. Nayak, "Improving recommendation novelty based on topic taxonomy," in *Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, (Washington, DC), pp. 115–118, IEEE Computer Society, 2007.
20. D. Billsus and M. J. Pazzani, "User modeling for adaptive news access," *User Modeling and User-Adapted Interaction*, vol. 10, no. 2–3, pp. 147–180, 2000.
21. C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proceedings of the 14th International Conference on World Wide Web*, (New York, NY), pp. 22–32, ACM, 2005.
22. G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

Springer
http://www.springer.com/978-3-642-13286-5

Music Recommendation and Discovery
The Long Tail, Long Fail, and Long Play in the Digital Music Space
Celma, Ò.
2010, XVI, 194 p., Hardcover
ISBN: 978-3-642-13286-5