COLLABORATIVE FILTERING

JOHN ASHLEY BURGOYNE · ID 260172518

MUMT 611 · PROF. ICHIRO FUJINAGA · 22 MARCH 2007

COLLABORATIVE FILTERING (CF) is a very popular technique, especially in commercial applications, for recommending products of some kind to clients. It has been used extensively for music recommendation systems, and chances are that most of the music-listening public has obtained at least one piece in their personal collection from the output of a CF system. It requires a large database of user data to work properly, when such data exists, it is not difficult to implement.

The motivation for collaborative filtering comes from the notion of the "long tail" (Anderson 2004; Anderson 2006), better known in statistical literature as a heavy tail. *Tail* in the case refers to the tail of a probability density function (PDF). In a heavy-tailed distribution, a disproportionate amount of probability mass is distributed far from the mean (relative to a Gaussian distribution). Music sales, for example, would have a heavy-tailed distribution. Almost all of the sales are of a relatively small number of "hits" while the remainder of recording labels' collections moulder in obscurity. This large number of unknown albums constitutes a long tail.

Collaborative filtering seeks to encourage clients to explore the long tail, which is profitable for clients, who get to explore products they might not otherwise know existed, and for businesses, who stand to earn a profit on otherwise dead inventory. In the typical arrangement, a business asks their cleints to rate as many of its products as possible. This rating matrix will necessarily be sparse, as most clients will not have come into contact with most products. As a consequence of its size and sparsity,

however, the matrix will also contain singularities which can be leveraged to simplify it, if sometimes only in theory, to a non-sparse matrix of user groups. Many techniques have ben used to accomplish this simplification [CITE].

Collaborative filtering can be done in a user-based or item-based form. The user-based form matches the description above: users rate every product, and the filtering process identifies users who have made similar ratings to the user requiring a recommendation. The idea it to combine the ratings of similar users to predict how any given user would rate products he or she has not seen. The item-based approach, in contrast, cross-correlates the items themselves, usually based on something like a purchasing history. The most famous example of item-based collaborative filtering could be the "Customers who bought this item also bought..." feature at `http://www.amazon.com`.

Besides Amazon.com and its competitor Barnes & Noble, other notable commercial applications of collaborative filtering include TiVo, the digital television recording and recommendation system,[1] and Netflix, a mail-order DVD rental programme that recommends new films to see based on order history. In the musical domain, successful applications of collaborative filtering have included the Musicmatch Jukebox (now part of Yahoo!) and Last.fm (`http://www.last.fm`). The Audioscrobbler plug-in records the listening habits of voluntary Last.fm subscribers and uses this information to generate custom radio stations. Like any collaborative filering system, as Last.fm's user base has grown, so has the success of its recommendations.

Active research questions remain open in the area, including applications to music. One of the most important questions being asked is how to integrate information form social systems like collaborative filtering with more traditional content-based recommendation (Berenzweig et al.

[1] TiVo's recommendations are not always, however, to their clients' liking. Read about one man's struggle to convince his TiVo of his orientation in Zaslow 2002.

2003; Yoshii et al. 2006). Although collaborative filtering has been less successful in generative complete playlists, it also has some uses in this domain (Cunningham, Bainbridge, and Falconer 2006). Finally, there are questions about how to generate enough data to make collaborative filtering useful in a research context. One promising such method is web mining (Cohen and Fan 2000).

REFERENCES

Anderson, C. 2004. The long tail. *Wired* 12 (10).

Anderson, C. 2006. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion.

Berenzweig, A., B. Logan, D. P. W. Ellis, and B. Whitman. 2003. A large-scale evaluation of acoustic and subjective similarity measures. In *Proceedings of the International Conference on Music Information Retrieval*.

Cohen, W. W., and W. Fan. 2000. Web-collaborative filtering: Recommendign music by crawling the Web. *Computer Networks 33*: 685–98.

Cunningham, S. J., D. Bainbridge, and A. Falconer. 2006. 'more of an art than a science': Supporting the creation of playlists and mixes. In *Proceedings of the International Conference on Music Information Retrieval*.

Yoshii, K., M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. 2006. Hybrid collaborative and content-based music reocmmendation using probabilistic model with latent user preferences. In *Proceedings of the International Conference on Music Information Retrieval*.

Zaslow, J. 2002, 26 November. If TiVo thinks you are gay, here's how to set it straight. *Wall Street Journal*.