CSC2231 – Internet Systems and Services

Paper Review – Kazaa
Name: Alex Wun
Date: Nov. 27th, 05

The authors of this paper obtain a roughly 6 to 7 month trace of Kazaa traffic from clients within the University of Washington. They first notice that Kazaa requests do not consist of a single type of workload - object sizes range dramatically from <10MB to nearly 1GB. And although the vast majority of requests are for small objects, most of the bytes transferred are due to large objects.

More importantly, the authors find that the popularity distribution is not zipf. Instead, they show that the distribution is similar to a *fetch-at-most-once* object popularity distribution. Effectively, the most popular 10% of the content (for >100MB objects) exhibits a "flattened" popularity in comparison to the zipf model. The justification given by the authors is that due to the time and bandwidth required to download these objects, users only request these objects once. Hence, the number of hits for each object is lower than predicted by zipf. I believe this is correct, but there is probably also the factor that I mentioned in my "long tail" review: the zipf distribution is partially an artifact of the indexing/searching mechanisms that users inherently go through to obtain content.

Fetch-at-most-once does not seem to be the only explanation since book and music CD popularity is supposedly zipf (implicitly from the "long tail" article) … but people do not buy the same book or CD repeatedly. So the very popular content is most likely due to *unique* "hits" for those books or CDs. Perhaps the flattened popularity distribution in Kazaa is a result of the searching scheme. There are no prerequisite "entry point" objects (home pages, well-known book/CD titles) that collect "useless" hits in Kazaa since users can search for and download the exact niche content they desire (provided it exists). This would indicate that the popularity distribution shown more closely reflects the "true" popularities of various multi-media objects. Another indication of this is the fact that the *fetch-at-most-once* curve is actually flat for the popular content. The Kazaa curve is not flat, but simply exhibits a slow but steady climb in popularity.

It would be interesting to model the objects of a web site as an n-ary tree – with the root representing the homepage, the first level children representing the site's main sections and so on. If the requests for the "leaf" node objects are uniformly distributed, what does the number of accesses to each node look like (given that a node is "accessed" when it is traversed to reach the desired leaf node) across multiple trees (web sites).