MS&E 235: Lecture 9
Prof. Ashish Goel
Apr 28, 2009
Notes by Rio Goodman

Lecture 9: Reputation Systems

Reputation systems exists in many areas of internet commerce, from hotel reviews to search results. There are systems that rank users, products, hotels, services and websites. Rankings are important on the internet because there is so much information available that it is useless if it is not ranked. For example, a Google search for "purple rain" returns 1.5 million results. If these results were not ordered in a helpful manner, it would take too long to look through all the webpages. If we look up a specific model of camera on epinions.com, we get a list of reviews. The reviews are the reputation system for the products being reviewed, but the reviews themselves are also ranked. The reviews are displayed in order of reviewer reputation, so there are two reputation systems here.

Dangers of reputation systems

With so many reputation systems, it is important to understand the dangers involved: bad mouthing, ballot stuffing, personalization, and presentation bias all lead to inaccurate rankings. Bad mouthing is when a competitor writes a number of bad reviews, bringing the reputation down. Ballot stuffing is when someone writes a number of good reviews for themselves (or pays someone to), bringing the reputation up. Personalization is a term for the fact that reputation should really vary depending on the user. For example, one search result may be more interesting to one user than to another. And presentation bias is the inherent bias that comes from the order in which things are displayed (ie. the first thing listed will have some advantage over the second thing). While these dangers also exist in interpersonal reputations, they are more pronounced online because of the automation, anonymity, and the easy availability of cheap pseudonyms. For example, it is easy for a hotel to write a lot of good reviews for itself. Different websites deal with these dangers in different ways, but most do not disclose their methods.

Online examples of reputation systems

Epinions.com allows people to express their opinions about all sorts of products. As mentioned above, it has two reputation systems, one for products and one for users.

Tripadvisor.com is a website that allows users to rate hotels (and restaurants, cruises, etc.) and write reviews. It also allows users to search for specific hotels or hotels in a region, and see what other users have said about them. This is an example of a reputation systems where users' feedback creates reputations for hotels. Like Epinions, Tripadvisor also has a reputation system for the order the reviews themselves are displayed. Websites such as Tripadvisor are highly susceptible to fraud, since a hotel can write a review on itself, or hire people to write a good review for it. Interestingly, Tripadvisor has no policy against this, and does not police the legitimacy of reviews.

EBay is home to another reputation system. Whenever a transaction occurs on eBay, the buyer and seller rate each other. This is highly influential, especially for new users. If a new seller is attempting to enter the market on eBay, one bad sale could tarnish his reputation and influence potential buyers negatively.

Imdb.com also has a very influential reputation system in its movie ratings. Many people see the ratings on imdb.com, and if a movie has a high rating, users are more likely to rent or purchase it.

Lastly, and most importantly, Google has a reputation system. The order in which search results are displayed is determined by the "PageRank" of the web pages that appear in the search[1]. Since Google receives far higher traffic than any of the other examples mentioned, its reputation system is very important.

PageRank

The order in which search results are displayed is perhaps the most common example of a reputation system. Even for obscure searches, there can still be many many search results. So how do search engines order all these results so that they are most useful for the user? The answer is: they order them by "reputation." One such measure of reputation is PageRank, which uses the hyperlinked nature of the web. In the original Google search engine, every web page was assigned a PageRank, and search results were then displayed in decreasing order of PageRank.

If we think of PageRank as an analogy to reputation among people, then if page A links to page B, then it is as if B is thought highly of by A. If many people think highly of B, then B's reputation increases, and the same is true of PageRank. But number of links is not the only determining factor. The reputations of the people that think highly of B also matter. If ten people think highly of B, but those ten people are not thought highly of by anyone else, then B's reputation should not be as high as if B's ten fans were thought highly of by others. From this conceptual idea of reputation, we can start to construct the mathematical definition of PageRank.

Think of the internet as a graph, $G(V, E)$, with a set of vertices, V, and a set of edges, E. Each vertex $v \in V$ represents a webpage. If there is a link from page v to page w, then there is a corresponding directed edge in the graph, $(v, w) \in E$. We let d_v denote the number of edges coming out of v (out degree), and π_v the PageRank of v. From our previous discussion of reputation, we can construct the following definition of PageRank.

$$\pi_v = \sum_{(u,v)\in E} \frac{\pi_u}{d_u}$$

We will assume the following constraints:

$$\pi_v \geq 0, \forall v$$

$$\sum_v \pi_v = 1$$

With this definition, we have captured the idea of reputation. If page A links to page B, A contributes a fraction $(1/d_A)$ of its PageRank to B. B collects the contributions from all the pages that link to it, to determine its own PageRank. This is **naïve PageRank**, and is equivalent to random surfing.

Random surfing can be thought of as a monkey surfing the internet. At every time period, the monkey randomly clicks a link on his current page. This means that the likelihood that the monkey clicks a certain link is one divided by the number of links on the page. If page A and C both have links to page B, then the fraction of time the monkey will spend at page B is (fraction of time he spends at A)×(likelihood he clicks on link to B while at A) + (fraction of time he spends at C)×(likelihood he clicks on link to B while at C). Thus we can construct the following expression for the fraction of time the monkey spends at a given page, v.

$$\pi_v = \sum_{(u,v)\in E} \frac{\pi_u}{d_u}$$

such that

$$\pi_v \geq 0, \forall v$$

$$\sum_v \pi_v = 1$$

This is the same expression as the definition of PageRank.

[1] In fact, Google currently uses undisclosed methods to order the search results, but the original Google method was PageRank, and the current method is likely based on PageRank.